Session 4 Overview: *Digital Processors*

DIGITAL ARCHITECTURES AND SYSTEMS SUBCOMMITTEE

Session Chair: *Mahesh Mehendale, Texas Instruments, Bangalore, India*

Session Co-Chair: *Luke Shin, Oracle, Santa Clara, CA*

Improved performance, battery-life and interactive user-experience continue to drive advances in digital processors for PCs, smartphones, tablets, ultra-light laptops and automobile infotainment systems, as described in the first four papers of this session. The fifth paper describes techniques to meet safety standards for automotive processing. The final two papers address ultra-low power applications, such as sensor node processing and wearables, using special-purpose acceleration hardware and non-volatile memory for zero power standby modes.

1:30 PM

4.1 14nm 6th-Generation Core Processor SoC with Low Power Consumption and Improved Performance

E. Fayneh, Intel, Haifa, Israel

In Paper 4.1, Intel presents its sixth-generation Core processor, a 14nm fully featured SoC, which provides comparable performance for less than 50% power versus a 22nm Ultrabook processor.

2:00 PM

4.2 Increasing the Performance of a 28nm x86-64 Microprocessor Through System Power Management

A. Grenat, AMD, Austin, TX

In Paper 4.2, AMD presents a range of system techniques to manage power, thermal, voltage margin, and reliability to increase the effective performance of a 28nm x86-64 microprocessor by 15% within the same process node.

2:30 PM

4.3 A 20nm 2.5GHz Ultra-Low-Power Tri-Cluster CPU Subsystem with Adaptive Power Allocation for Optimal Mobile SoC Performance

S. Gururajarao, MediaTek, Austin, TX

In Paper 4.3, MediaTek presents industry's first tri-cluster, 10-core CPU, featuring three ARMv8a CPU clusters optimized for 1.4GHz, 2.0GHz, and 2.5GHz operation in a 20nm high-κ metal-gate process. Compared to dual-cluster CPUs, the addition of a third cluster provides 40% higher overall performance with 40% improved power efficiency.

 978-1-4673-9467-3/16/$31.00

3:15 PM

4.4 A 197mW 70ms-Latency Full-HD 12-Channel Video-Processing SoC for Car Information Systems

S. Mochizuki, Renesas System Design, Tokyo, Japan

In Paper 4.4, Renesas presents a 12-channel HD-video-processing SoC, implemented in 16nm FinFET CMOS, with 197mW power consumption and 70ms video latency for automobile infotainment and driver-assistance applications.

3:45 PM

4.5 A 16nm FinFET Heterogeneous Nona-Core SoC Complying with ISO26262 ASIL-B: Achieving 10^{-7} Random Hardware Failures per Hour Reliability

C. Takahashi, Renesas System Design, Tokyo, Japan

In Paper 4.5, Renesas presents a heterogeneous nine-core SoC complying with the ISO26262 ASIL-B standard for automobile safety, achieving a reliability rate of less than 10^{-7} random hardware failures per hour, in the same 16nm FinFET SoC as described in paper 4.4.

4:15 PM

4.6 A 65nm CMOS 6.4-to-29.2pJ/FLOP@0.8V Shared Logarithmic Floating Point Unit for Acceleration of Nonlinear Function Kernels in a Tightly Coupled Processor Cluster

M. Gautschi, ETH Zurich, Zurich, Switzerland

In Paper 4.6, ETH Zurich presents a 6.4-to-29.2pJ/FLOP (at 0.8V supply) tightly coupled quad-core processor cluster in 65nm CMOS, with a shared-logarithmic floating-point unit for acceleration of nonlinear function kernels.

4:45 PM

4.7 A 65nm ReRAM-Enabled Nonvolatile Processor with 6× Reduction in Restore Time and 4× Higher Clock Frequency Using Adaptive Data Retention and Self-Write-Termination Nonvolatile Logic

Y. Liu, Tsinghua University, Beijing, China

In Paper 4.7, Tsinghua University presents the first ReRAM-enabled nonvolatile 100MHz micro-controller in a 65nm CMOS-logic-compatible process, with 20ns restore time and 0.45nJ restore energy.

4.1 14nm 6th-Generation Core Processor SoC with Low Power Consumption and Improved Performance

Eyal Fayneh, Marcelo Yuffe, Ernest Knoll, Michael Zelikson, Muhammad Abozaed, Yair Talker, Ziv Shmuely, Saher Abu Rahme

Intel, Haifa, Israel

Intel's 6th generation Core processor (code named "Skylake" or SKL) was designed to enable PC performance and user-experience at smaller and thinner form factors and enable fan-less PC platforms. It required optimization to an extremely low thermal design point (TDP). The lower average power consumption of SKL vs. the previous generation considerably increases the system battery life and allows full-day battery life or thinner designs with smaller batteries. The SKL product family is manufactured using an Intel 14nm tri-gate CMOS 11-metal-layer technology, as with the previous Core generation. Different dice include: 2 or 4 cores, a shared last-level cache (LLC, 1MB/core), a scalable graphic processor (GP) with 24, 48 or 72 execution units (EU), an image processing unit (IPU, supporting 4 cameras simultaneously), 2 channels of DDR3/LPDDR3/DDR4, a display engine (DE) and 3 display IO ports configurable to eDP, DP or HDMI. In mobile SKUs, the peripheral control hub (PCH) resides in the same package (MCP) as the CPU and communicates through an on-package IO (OPIO) bus. For desktop (DT), the PCH resides on the platform. Fig. 4.1.1 presents the SKL block diagram for the minimum configuration (2 cores, 24 EU GP, MCP). A key challenge was the need to add new capabilities, while reducing power, especially for some of the popular uses (media, casual gaming, speech recognition and advanced imaging).

Figure 4.1.2 shows the summary of the SKL post Si power consumption in different usage scenarios. All uses fit into tablet form factors, and all (except 4K video capture) can fit into 5.8" high-end smartphone form factors. Special attention was given to 4K panels, that are gaining momentum as the premium PC display standard, while 4K content is aggressively ramping. Power scaling from a 28×18 panel to a 4K panel is in the range of 1.1-1.3× for a 1.7× pixel increase.

Figure 4.1.3 shows SKL post Si power vs. a 14nm previous generation [1] comparable processor. At video playback (VPB), the power reduction is 45%: 21% by general HW optimization and 24% by specific VPB optimization (supporting NV12 format in the DE, frames scaling in the DE instead of the GP, display buffer size optimization and a dedicated fixed function for video decode in the GP). Fig. 4.1.4 shows the SoC power improvement impact on total system battery-life. New power technologies: burst operation, multi-power-plane (MPP), local power gating (PG), local data retention, high-granularity clock gating, clock domains consolidation, lower V_{min}, DVFS for the system agent (SA), memory compression features, low-power PLLs and IO power optimization enabled wider dynamic ranges. The low-power version of the product is designed to operate fan-less at sustained 4.5W, but can burst up to 15W for short durations to deliver responsiveness and full PC experience. As a result, SKL performance at 7W TDP is comparable to a 15W Ultrabook 2 years back [3].

The power saving features led to a wide power dynamic range of 3 decades (peak power to connected stand by). Supporting such a wide power dynamic range, while minimizing the processor area imposes a big challenge for the power delivery network (PDN) design. Extensive use of MPP logic, local PG and local data retention using dedicated always-on power supplies are examples of the architecture and design solutions devised to deal with this challenge. In order to support communications among geographically disjoint, multiple-power-domain areas of an SKL die, local fine-grained power regions were enabled. Implementation of local power islands was done using dual power grids on the section-level metals and non-gridded power routing on the top metals, which resistivity enables support of local domains without utilizing package routing. Different power domains were analyzed using industry standard PDN pre-silicon simulators and factoring in local timing criticality, in order to define the sufficient amount of metal resources for the PDN. PDN infrastructure enables up to 34 domains: one separately gated functional block in each IA core; 10 independent PDN sections in the SA; 8 separate power gated regions in DDR-IO. PG turn-on time was defined based on the design tolerance to the associated latency. In the SA typical turn-on time is ~150ns, while in the IA cores, the local gated region needs to become operational within 20ns. As local power gate toggling is not correlated with operation of surrounding logic, noise induced by ON↔OFF transitions in the victim neighborhood circuits should not exceed a crosstalk limit during normal operation. Another parameter that must be factored is reliability stress that develops when turn-on current exceeds the allowed maximum power density. Dedicated analog circuitry that enables simultaneous power-up of all power gate transistors in a controlled and tunable pace avoids reliability issues. In-situ retention of architectural registers using a low-power (~1mW) always-on power supply was enabled using dedicated lumped and distributed power supply multiplexers.

The power required to generate and distribute the SoC clocks imposes other important challenges in advanced SoCs. In SKL, all the clocks are generated by digital PLLs that were designed for performance at low power, saving 60% power compared to compared to previous-generation analog PLL (7mW at 4GHz frequency). The display IO (DP, eDP, and HDMI) is clocked by fractional-N (Fn) divider-less digital PLLs, one dedicated PLL for each display IO port. The Fn-PLL permits generation of accurate (<1PPM) clock frequency, avoiding the need for an external phase-interpolator that was required to modulate the reference signal to achieve the 1PPM frequency accuracy in previous generations. The SSC (spread-spectrum-clock) modulator is implemented within the PLL avoiding the use of 2 reference signals for SSC and non-SSC applications (DP-SSC vs. HDMI-non-SSC), reducing platform power. The PLL time-to-digital converter (TDC) circuit gates the DCO clock to ±5% of the reference clock period surrounding the reference falling edge (Fig. 4.1.5) to reduce the TDC power from 10mW to 1mW. The PLL fine accuracy allows generation of any transfer-rate required by the display resolution and protocol, requiring fewer IO lanes, thus increasing the power efficiency of the display block. The Fn-PLL saves 17mW@eDP mode and 44mW@HDMI mode vs. the previous generation.

The clock is distributed in each of the logic domains by a structure of spines [2] and clock islands. This approach enables the implementation of multiple gated clock derivatives for power conservation. For example, the core clock distribution (Fig. 4.1.6) was split into 58 different logical clock domains with different enable signals to control the activity factor of each domain. Activating the clock-gating feature reduces the clock dynamic capacitance (Cdyn) at TDP conditions by 26% (up to 2.3% less IA core total Cdyn, which translates into additional performance gain for power-limited systems). In the SA area, some clock domains were merged to optimize power. The clock for these blocks is generated by a single PLL driving multiple clock spines or driving multiple clock and voltage domains from a shared clock spine. The required clock frequency division and voltage level shifting is done at the clock global driver outputs. This saved both area and power by reducing the number of PLLs and clock spines.

Several techniques were used to optimize the power in the analog front-end (IO circuits). Reducing pad capacitance (C_{PAD}), while still meeting the R_{ON} design target saves power for both reads and writes according to the following formula: $\Delta Power = \Delta Cpad \bullet V_{swing}^2 \bullet freq \bullet \#pins \bullet AF$ (AF is the bus activity factor). Second order power savings can be achieved by further optimization of the output buffer pre-driver circuitry. The delay-line loops (DLL) that are used in the DDR PHY have a relatively high activity and therefore are one of the main power consumers – they became the main target of the power reduction efforts. Active power reduction was implemented by cutting the delay-line length, while keeping them locked on the same T_{CYCLE} delay. This reduces the active power of the delay line by the amount of capacitance to be charged within the same clock cycle. To reduce idle power, the DLLs were set to power-saving mode when there are no read or write transactions on the bus. SKL DDR IO implements power domain partitioning. This is a new approach in which all the purely digital logic was moved to a dynamically regulated power domain based on the CPU operating point and internal analog blocks were moved to a lower voltage domain.

Figure 4.1.7 shows a photo of the die; several different versions of SKL were introduced into the market in Q3/2015; these different versions allow a variety of system-level power/performance optimization points to cover the full spectrum of the demanding PC market segments.

References:

[1] A. Nalamalpu et al., "Broadwell: A Family of IA 14nm Processors," *Symp. VLSI Circuits*, pp. C314-C315, 2015.

[2] M. Yuffe et al., "A fully Integrated Multi-CPU, GPU and Memory Controller 32nm Processor," *ISSCC Dig. Tech. Papers*, pp. 264-265, 2011.

[3] N.Kurd et al., "Haswell: A Family of IA 22nm Processors," *ISSCC Dig. Tech. Papers*, pp. 112-113, 2014.

 978-1-4673-9467-3/16/$31.00

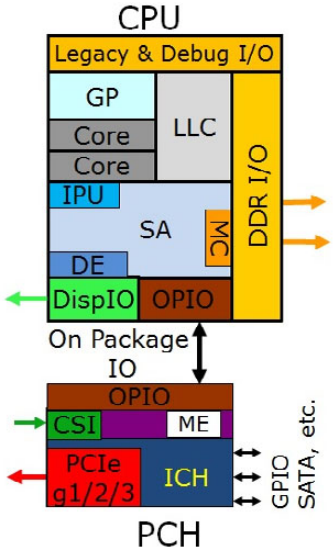


Figure 4.1.1: SKL block diagram.

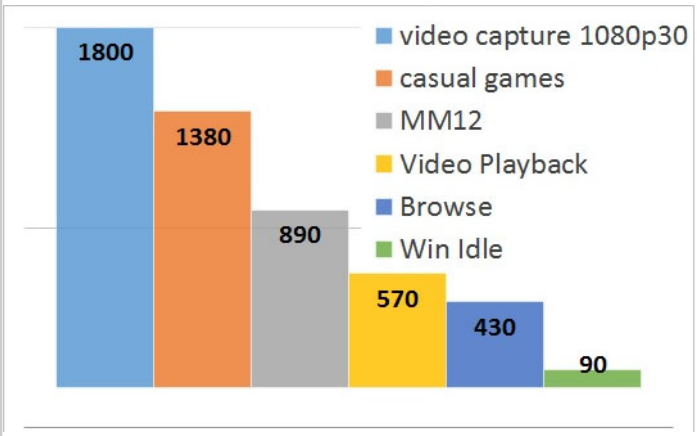


Figure 4.1.2: SKL usage scenario power [mW].

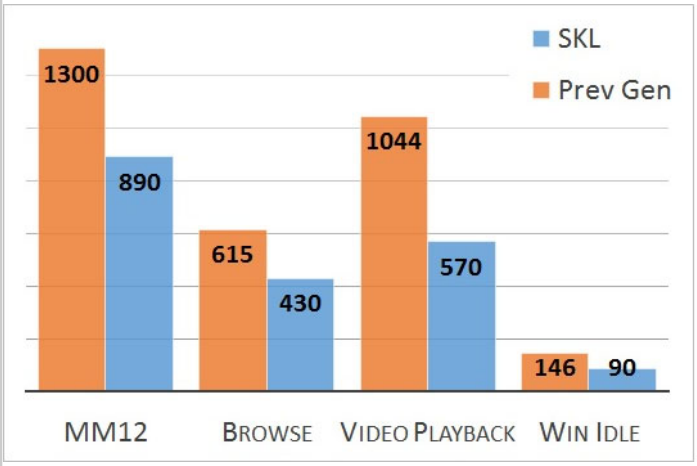


Figure 4.1.3: Power of SKL vs. a 14nm previous generation for different usage scenarios [mW].

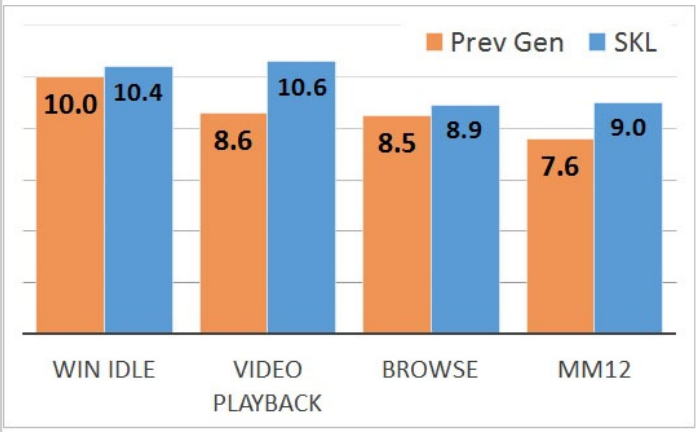


Figure 4.1.4: SKL usage scenario battery life [hours].

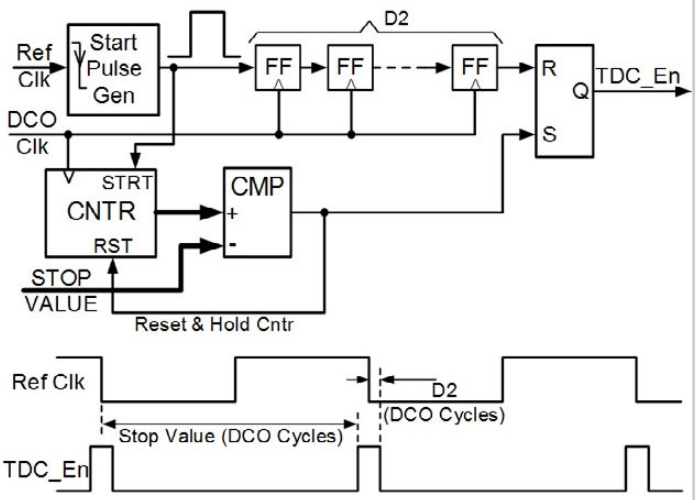


Figure 4.1.5: Display PLL TDC gating circuit.

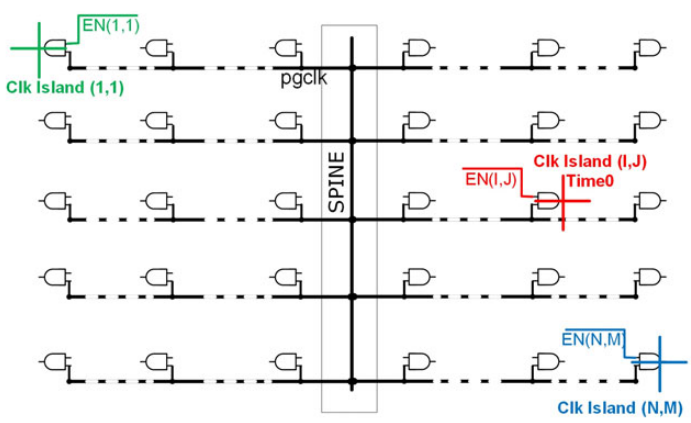


Figure 4.1.6: SKL core clock distribution concept block diagram.

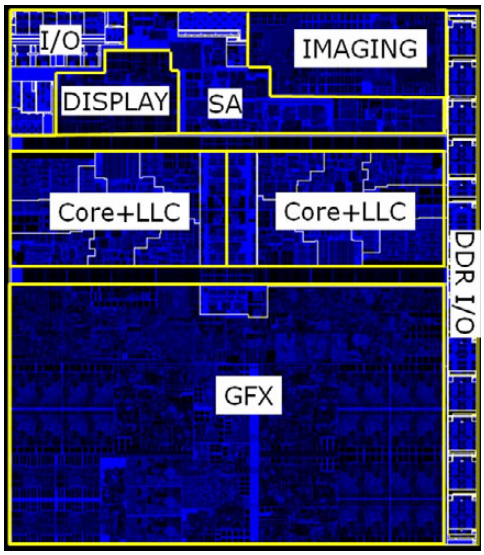


Figure 4.1.7: SKL die photo.

978-1-4673-9467-3/16/$31.00

4.2 Increasing the Performance of a 28nm x86-64 Microprocessor Through System Power Management

Aaron Grenat[1], Sriram Sundaram[1], Stephen Kosonocky[2], Ravinder Rachala[1], Sriram Sambamurthy[1], Steven Liepe[2], Miguel Rodriguez[2], Tom Burd[3], Adam Clark[4], Michael Austin[1], Samuel Naffziger[2]

[1]AMD, Austin, TX,
[2]AMD, Fort Collins, CO,
[3]AMD, Sunnyvale, CA,
[4]AMD, Markham, ON, Canada

Power-management techniques can be effective at squeezing more performance and energy efficiency out of mature SoCs. V_{max} reliability limits, infrastructure limits, guard-bands, aging, and thermal limits all put restrictions on performance. This paper describes five power-management techniques that provide a net performance increase of up to 15%, depending on the application and TDP of the SoC, on "Bristol Ridge", a 28nm CMOS dual-core x86 APU.

Reliability Tracker: In a high-performance microprocessor, the critical limiters to long-term reliability are time-dependent dielectric breakdown (TDDB) and electromigration (EM). TDDB is enormously dependent on voltage, and both TDDB and EM are strongly dependent on temperature. Standard microprocessor design assumes worst-case conditions on voltage and temperature for long-term reliability targets, suppressing potential performance when operating in more typical conditions.

Bristol Ridge utilizes a reliability tracker to dynamically monitor operating conditions, voltage (V_{dd}) and temperature (T), for each compute unit (CU), and estimates an aggregate failure, or FIT rate, for TDDB and EM in 1ms increments: $FIT = k_{TDDB}e^{b_{TDDB}T}V_{dd}^{\gamma_{TDDB}} + k_{EM}e^{b_{EM}T}V_{dd}^{\gamma_{EM}}$, where the subscripted variables are curve-fit from foundry process reliability models. The FIT rate is a function of voltage and temperature as shown in Fig. 4.2.1. The FIT rate is calculated in 1ms increments, faster than the frequency of P-State changes, so that only one V_{dd} and temperature measurement is required per interval. The results are then filtered to prevent fast P-State dithering: $FIT\_filtered_N = \alpha \bullet FIT + (1-\alpha) \bullet FIT\_filtered_{N-1}$. The overwhelming majority of users can realize a 100MHz frequency boost, while staying within long-term reliability targets.

Digital LDO (DLDO): Fig. 4.2.2 shows the architecture of the fully digital on-die voltage regulation system (DVS). The system employs a PSM (power-supply-monitor) [1] to sample the core voltage in a feedback loop that modulates the power header resistance across input voltage and V_{reg}. We repurpose part of the power gating headers as a digitally controlled series resistance to drop voltage from a shared input with another core-pair in the SoC, at a very small cost to the core area. This system is designed as a simple voltage regulator, without the need for additional output capacitance sufficient for low-speed operation to primarily service cache probe operations.

To maintain simplicity, the DLDO operates on a 100MHz clock limiting its response time to large load current variations. In an effort to avoid adding costly capacitance (e.g., package capacitance) on the regulated node, we employed novel architectural di/dt throttling techniques to achieve a low-power state (P_{min}) using the DLDO. These techniques include "single issue operation", "disable branch prediction" and "non-speculation operation". In P_{min}, the core can safely operate at 800MHz/0.95V with a worst-case current transient workload (Fig. 4.2.3).

The "C-state" boost feature is a mechanism which allows active cores to exceed the normal P1 (configurable) P-State to a P0 boost P-State if enough of the cores are in a low-power (e.g., CC6) state. "C-state boost" is a major performance contributor for lightly-threaded workloads. P_{min} can be used to provide similar power as CC6 but faster exit latency and cache/core state retention, reducing idle transition latency by removing the need to flush the cache. The DLDO enables maximum residency in P_{min}, raising performance of lightly-threaded workloads since more time is spent in the "C-state boost", resulting in a net performance gain of ~6% in these scenarios.

STAPM (skin-temperature-aware power management): STAPM improves performance by boosting APU CPU and/or GPU frequencies for as long as the estimated platform skin temperature remains below the specified limit. This can result in significant performance improvements for some benchmarks, enhanced user-experience, and even improved battery life for race-to-idle computations. Bristol Ridge STAPM models the thermal capacitance of the platform using a simple alpha filter applied to the calculated APU power. This alpha filtered APU power is proportional to skin temperature. Alpha filter coefficients can be optimized by characterizing the chassis skin temperature response to APU power.

In mission-mode, skin temperature is managed by controlling APU power with a Proportional + Derivative + Accumulator algorithm in system management firmware. The inputs are an APU power limit that is proportional to the skin temperature limit and the alpha-filtered APU power. The output is a sequence of P-State assignments whose residency profile results in an effective CPU or GPU frequency. Fig. 4.2.4 shows the controller and results.

Boot-time calibration (BTC): Traditional binning includes fixed voltage margin to cover for voltage regulator module (VRM) DC tolerance and package variations. However, most reference systems offer tighter voltage delivery than worst-case specification. The voltage level is tracked in Bristol Ridge using on-die PSMs [1]. By comparing the PSM voltage in a quiescent condition during system "boot", with those on ATE (automated test equipment), BTC can accurately calibrate the DC component of the power supply difference. In order to tune out the temperature dependence, BTC is carried out at the "temperature inversion" voltage. This voltage represents the point at which transistor delay is virtually insensitive to temperature. This allows the BTC mechanism to accurately gauge the system DC loss.

The AC component of the voltage droop on a platform is quite different when compared to ATE due to power delivery and workload differences. For instance, ATE voltage droop at highest operating F_{max} is about 50mV, while in systems it is closer to 200mV. BTC employs a smart platform-specific calibration of per-part AC voltage droop as a function of frequency, temperature and per-part leakage, switching capacitance and droop measured on ATE. These DC and AC voltage droop components, as measured by BTC, allow the removal of worst-case margins. Fig. 4.2.5 shows how the voltages are aligned. Voltage savings from BTC were measured to be ~30mV, on average.

Moreover, BTC allows for tracking of per-part aging behavior. Traditionally, binning adds a worst-case margin to cover for aging (HCI/xBTI) effects over the target usage profile of the product. In contrast, since BTC now captures the PSM count of the part at each boot cycle and compares it to a fixed reference count measured originally on ATE, we find that the voltage margin added increases naturally as the part ages. During HTOL data collection, we prove the ring-oscillator within the PSM circuit aging compares very well with the actual critical path aging on the chip (Fig. 4.2.6). As a result, we can reduce the explicit aging guardband resulting in ~20mV of additional savings.

Shadow P-States**:** A critical path accumulator (CPA)-based scheme to accurately assess true Si speed capability and address the problem of voltage margin reduction in traditional binning flows was briefly reported in [2]. In the Bristol Ridge implementation of AVFS, we layer another capability: increasing peak frequency directly when headroom is available. AVFS allows us to exactly characterize the part-specific F_{max} capability, and BTC allows us to characterize the platform-specific power delivery margin. So combining AVFS and BTC we assess, at boot time, the peak frequency feasible for a given part in a given platform that meets the infrastructure limits (electrical design current and process V_{max}). We refer to these peak boost frequencies, as shadow P-States. In Bristol Ridge, shadow P-States enable peak boost frequencies, on average, to increase by 100MHz over conservative traditional binning.

References:

[1] K. Wilcox et al., "Steamroller Module and Adaptive Clocking System in 28 nm CMOS," *IEEE J. of Solid-State Circuits*, vol. 50, no. 1, pp. 24-34, 2015.
[2] K. Wilcox et al., "A 28nm x86 APU Optimized for Power and Area Efficiency," *ISSCC Dig. Tech. Papers,* pp. 84-85, 2015.
[3] Z. Toprak-Deniz et al., "Distributed System of Digitally Controlled Microregulators Enabling Per-Core DVFS for POWER™ Microprocessor," *ISSCC Dig. Tech Papers*, pp. 98-99, 2014.
[4] M. Saint-Laurent et al., "A 28nm DSP Power by an On-Chip LDO for High-Performance and Energy-Efficient Mobile Applications," *ISSCC Dig. Tech Papers*, pp. 176-177, 2014.

 978-1-4673-9467-3/16/$31.00

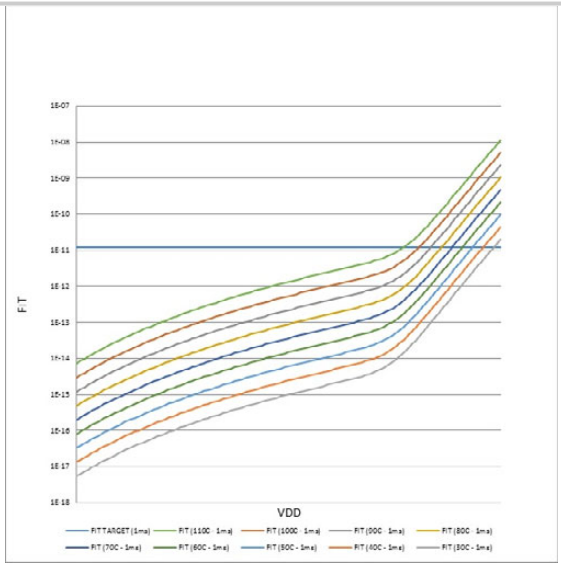


Figure 4.2.1: The FIT has a function of voltage and temperature and FIT target.

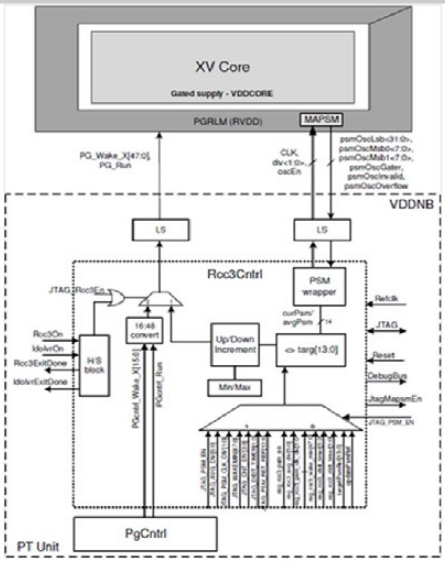


Figure 4.2.2: LDO controller block diagram.

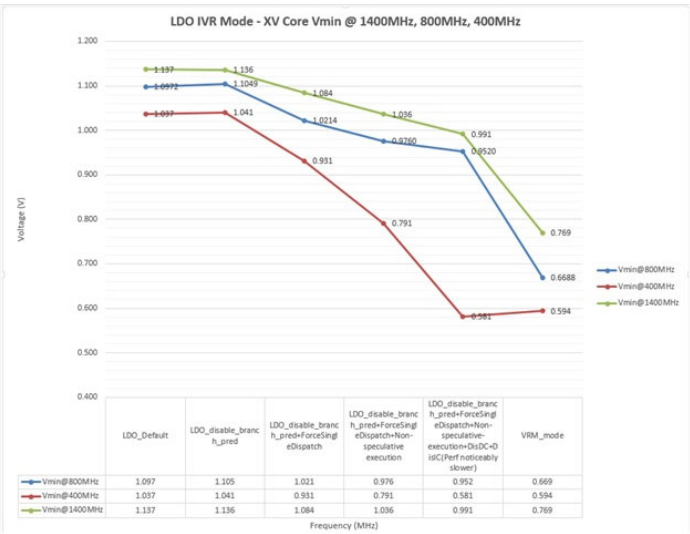


Figure 4.2.3: LDO results.

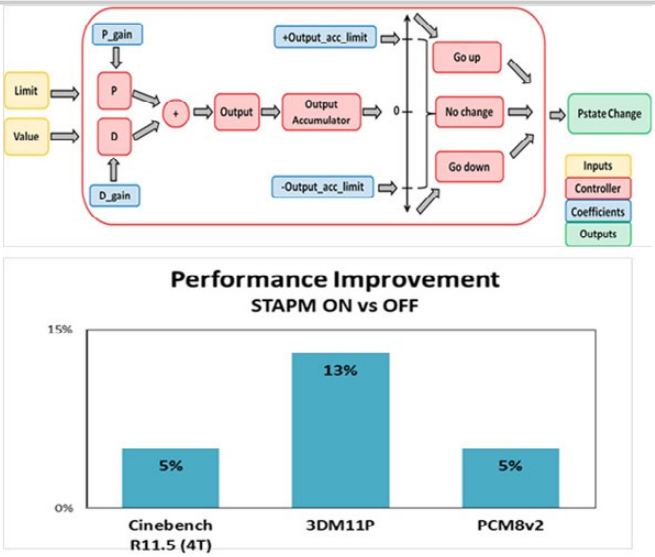


Figure 4.2.4: STAPM controller and benchmark results.

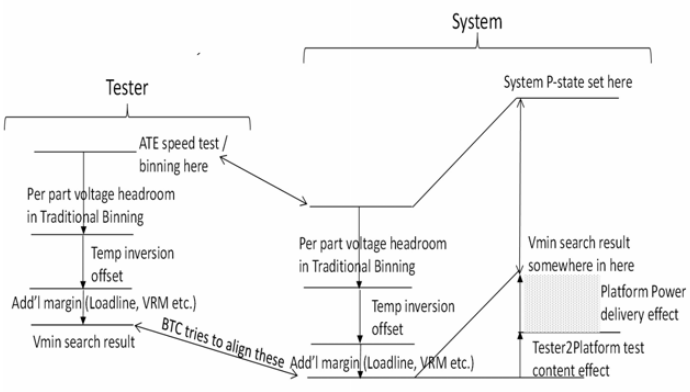


Figure 4.2.5: Aligning test voltage to a real system.

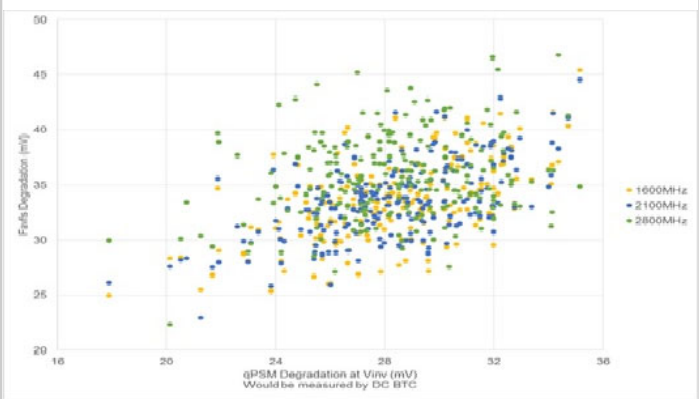


Figure 4.2.6: BTC estimated aging from PSM counts (X) matches actual paths aging (Y) across large operating range.

4.3 A 20nm 2.5GHz Ultra-Low-Power Tri-Cluster CPU Subsystem with Adaptive Power Allocation for Optimal Mobile SoC Performance

Hugh T Mair[1], Gordon Gammie[1], Alice Wang[2], Rolf Lagerquist[1], C.J. Chung[1], Sumanth Gururajarao[1], Ping Kao[2], Anand Rajagopalan[1], Anirban Saha[3], Amit Jain[4], Ericbill Wang[2], Shichin Ouyang[5], Huajun Wen[1], Achuta Thippana[1], HsinChen Chen[1], Syed Rahman[1], Minh Chau[1], Anshul Varma[1], Brian Flachs[1], Mark Peng[2], Alfred Tsai[2], Vincent Lin[2], Ue Fu[2], Wuan Kuo[2], Lee-Kee Yong[2], Clavin Peng[2], Leo Shieh[2], Jengding Wu[2], Uming Ko[2]

[1]MediaTek, Austin, TX,
[2]MediaTek, Hsinchu, Taiwan,
[3]MediaTek, Singapore, Singapore,
[4]MediaTek, Bangalore, India,
[5]MediaTek, San Jose, CA

This paper describes design features of the high-performance CPU from a heterogeneous tri-cluster, deca-core CPU subsystem incorporated into the Helio X20 mobile SoC for smartphone applications. The SoC is fabricated in a 20nm high-κ metal-gate CMOS, and has a die size of 100mm^2. Additional key features of the SoC include: a graphics processor unit, multimedia (including 32MPixel/24fps camera support), and connectivity subsystems integrating 802.11ac, GPS, and multistandard cellular modems, featuring LTE FTD/TDD R11 Cat-6 with 20+20 carrier aggregation (300/50Mb/s) DC-HSPA+, TD-SCDMA, Edge, CDMA2000 1x/EVDO Rev. A (SRLTE).

As shown in Fig. 4.3.1, the deca-core compute function contains three separate clusters of ARMv8a CPUs. A first cluster contains four power-efficient Cortex-A53 cores optimized for ultra-low power (ULP) applications, while achieving a maximum frequency of 1.4GHz. A second cluster contains four Cortex-A53 cores optimized for higher performance, at 2GHz, but also maintaining low power (LP). While the LP cluster has a higher power/operation than the ULP cluster, the LP cluster still maintains a power efficiency advantage over a third high-performance (HP) cluster, which contains two Cortex-A72 cores, operating at a maximum frequency of 2.5GHz, featuring out-of-order execution, and a 1MB L2 cache. Heterogeneous multi-processing (HMP) is further extended from quad-core [1], octa-core [2][3], to the tri-cluster, deca-core CPU subsystem with automatic adjustment of CPU resources according to the system workload. A die photograph (Fig. 4.3.7) highlights the three clusters.

A plot of power vs. single-thread CPU performance for all clusters is shown in Fig. 4.3.2. In contrast to 2-cluster approaches, the newly introduced LP cluster extends ULP-CPU power-efficiency benefits upward by 40% in performance (Fig. 4.3.2). An enhanced HMP cluster migration mechanism for tri-cluster balances performance and power for optimal adaptation to different system workloads. Moreover, the LP cluster achieves 40% better power efficiency for applications requiring a performance level that previously can only be fulfilled by the HP cluster in a dual-cluster CPU computing subsystem. Power/frequency optimization targets for the three clusters are designed to provide a continuous operation for a 10× dynamic performance range, but only at up to 4× power difference.

Adaptive power allocation (APA) is used to maximize CPU performance within the currently allocated power budget by instantly re-allocating power from low-activity CPUs to high-activity CPUs, thus avoiding performance throttling on high-activity CPUs. In scenarios where the cumulative power of all CPUs exceeds the total cluster budget, automatic clock gating is introduced as a temporary counter-measure and is achieved by clock-dithering 'APA-CD'. When APA-CD is active, a secondary process adjusts the on-chip PLL frequency and off-chip DC-to-DC converter voltage to a more energy-efficient operating point in order to maximize performance (MP); this process is called APA-MP. APA-CD and -MP interact to achieve maximum performance, which occurs when total power consumption is at the limit of the allocated budget with APA-CD disabled. Both APA-CD and -MP control loops have configurable digital filtering to adjust response rate and accuracy. Fig. 4.3.5 shows silicon results comparing the operating of APA-CD only vs. APA-MP+APA-CD both enabled – a >2× performance improvement is observed when power is limited to 25% of normal operating power.

Two new circuits are introduced to facilitate on-die power metering: a clock activity adder (CAA) for switching power, and a leakage monitor circuit for static power. Previously published approaches to switching power estimation involve monitoring key logic signals, which correlate to architectural and micro-architectural events and then applying scaling coefficients to each signal [4]. The approach of the CAA takes advantage of the fact that switching power is highly correlated to register clock activity, thus directly monitors, sums and accumulates the clock signals at the register pins (Fig. 4.3.3). Logically summing clock signals has the unique advantage of being able to comprehend the full clock hierarchy without monitoring upper-level clock-gate enables. As shown, if a clock is gated by a higher-level clock-gate, the leaf-level clock will also be gated, regardless of whether the leaf-level clock-gate itself is enabled or disabled. The correlation between metered power and simulated power is significantly improved with the CAA as shown in Fig. 4.3.3.

The leakage monitor operates by utilizing an exponential current DAC to source a leakage replica circuit (Fig. 4.3.4). The DAC is driven from a free-running gray-coded up/down counter. The direction of the counter is determined based on a comparison between the leakage replica voltage and the operating logic voltage, and converges to a ratio replica of the CPU leakage. By tracking the voltage on the operating logic, the leakage is automatically measured in active, retention, and off states without the need for additional power state monitoring.

Power delivery networks (PDN) are a critical design challenge in modern SoCs, and as a result, the ability to measure the performance of the PDN in end-equipment is a key value proposition. To achieve this an on-die high-bandwidth oscilloscope, SupplEyeScan (SES), is developed (Fig. 4.3.6). SES uses a 6b R2R DAC and a modified StrongARM latch [5] to sample a comparison of die-internal supply voltage and DAC output voltage. The DAC R2R ladder is supplied by externally applied 1.8V, and has an output range of 0.6-to-1.2V with ~10mV resolution. The StrongARM latch runs up to CPU clock speed, resulting in a measurement bandwidth of >1GHz. SES has three modes of operation: In histogram mode, up to n samples are recorded, where n is 255. These samples are spaced by up to m clock cycles, where m is 65535; sample spacing enables capture of low-frequency supply variation. By repeating the measurement and adjusting the DAC code, a voltage histogram is created. Once histogram limits are established, peak-detect mode is used to detect momentary excursions beyond the histogram range. By selecting either peak-high or peak-low, samples are taken continuously for n periods of m cycles, while recording any 'one' sample (for peak-high) or 'zero' sample (for peak-low) occurrence within each window of m cycles. Since $n \bullet m$ is up to 16M samples, the tails of the distributions are detected for a true limit of on-die voltage range. In transient mode, a software sequence is designed which stresses di/dt of the CPU load; an instruction within the software sequence issues a trigger signal to SES. By using the histogram method above, coupled with an increasing walk-off in the trigger delay, a cycle-by-cycle transient waveform is assembled to record the die-level PDN response to the load stress. A silicon-measured waveform is shown in Fig. 4.3.6.

In summary, a tri-cluster CPU complex is presented, which introduces a new LP cluster delivering a 40% increase in performance. Compared to DVFS of the HP cluster, the new LP cluster also achieves 40% improvement in power efficiency. Circuits are demonstrated which further increase performance in power limited scenarios through hardware-enforced run-time power limits, and a >2× performance improvement is achieved with APA-MP when the power budget is limited to 25% of normal capacity. Finally, performance of the power delivery network is validated in end-equipment through the integration of a high-bandwidth on-die oscilloscope.

References:

[1] A. Wang et al., "Heterogeneous Multi-Processing Quad-Core CPU and Dual-GPU Design for Optimal Performance, Power, and Thermal Tradeoffs in a 28nm Mobile Application Processor," *ISSCC Dig. Tech. Papers*, pp. 180-181, 2014.
[2] H. Mair et al., "A Highly Integrated Smartphone SoC Featuring a 2.5GHz Octa-Core CPU with Advanced High-Performance and Low-Power Techniques," *ISSCC Dig. Tech. Papers*, pp. 424-425, 2015.
[3] Y. Shin et al., "28nm High-k Metal Gate Heterogeneous Quad-Core CPUs for High-performance and Energy-efficient Mobile Application Processor," *ISSCC Dig. Tech. Papers*, pp. 154-155, 2013.
[4] V. Krishnaswamy et al. "Fine-Grained Adaptive Power Management of the SPARC M7 Processor," *ISSCC Dig. Tech. Papers*, pp. 74-75, 2015.
[5] Y. T. Wang and B. Razavi, "An 8-bit 150-MHz CMOS A/D Converter," *IEEE J. Solid-State Circuits*, vol. 35, pp. 308–317, Mar. 2000.

 978-1-4673-9467-3/16/$31.00

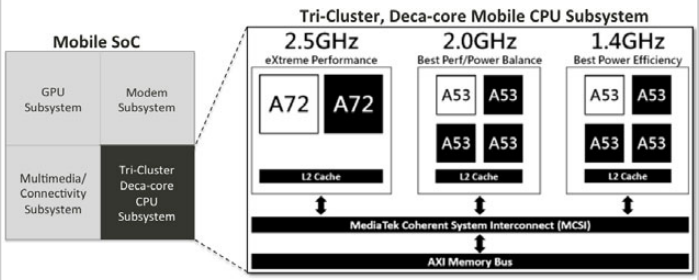


Figure 4.3.1: Mobile SoC and tri-cluster CPU subsystem architecture.

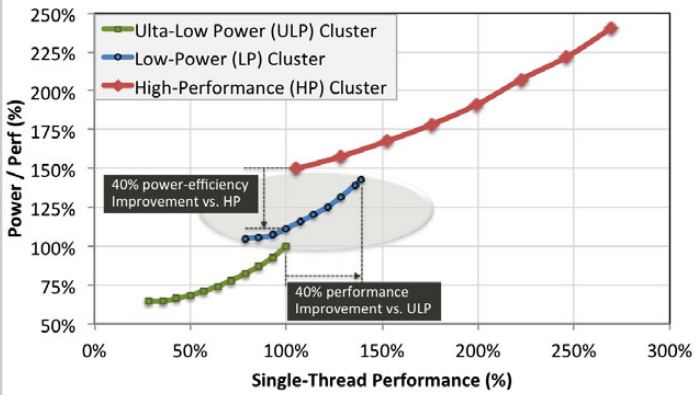


Figure 4.3.2: Tri-cluster power efficiency vs. performance dynamic range.

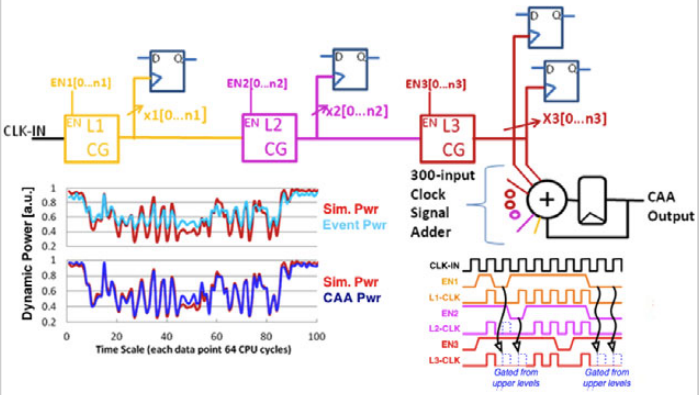


Figure 4.3.3: Clock activity adder and event accumulator.

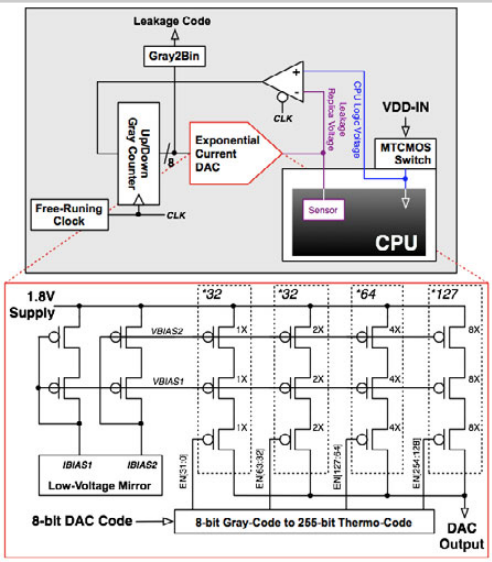


Figure 4.3.4: Leakage monitor block diagram and DAC schematic.

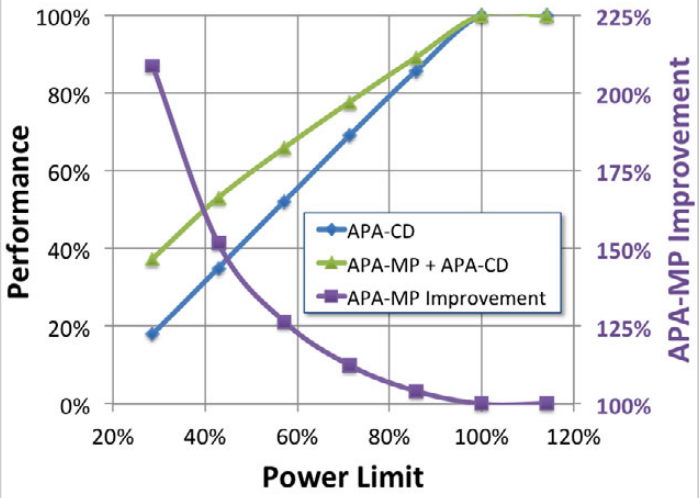


Figure 4.3.5: APA-MP performance benefit measure on silicon.

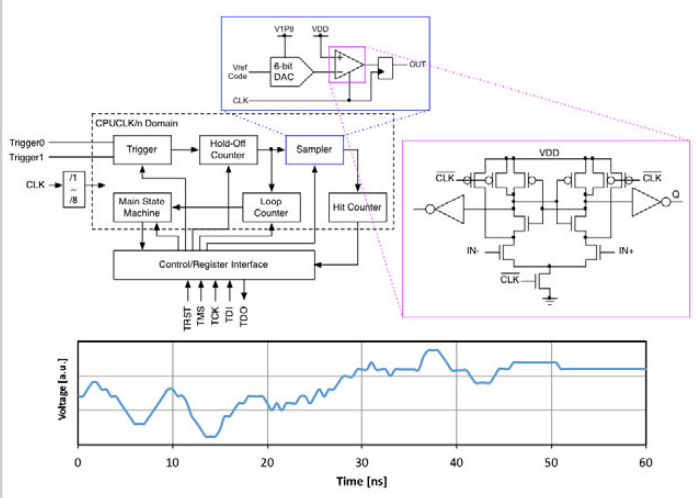


Figure 4.3.6: SupplEyeScan block diagram and silicon measured waveform.

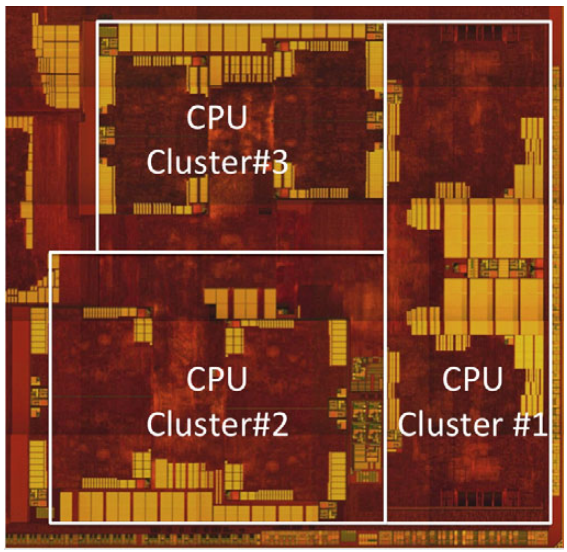


Figure 4.3.7: Die photograph of tri-cluster CPU subsystem.

978-1-4673-9467-3/16/$31.00

4.4 A 197mW 70ms-Latency Full-HD 12-Channel Video-Processing SoC for Car Information Systems

Seiji Mochizuki[1], Katsushige Matsubara[1], Keisuke Matsumoto[1], Chi Lan Phuong Nguyen[2], Tetsuya Shibayama[1], Kenichi Iwata[1], Katsuya Mizumoto[1], Takahiro Irita[3], Hirotaka Hara[3], Toshihiro Hattori[1]

[1]Renesas System Design, Tokyo, Japan,
[2]Renesas Design Vietnam, Ho Chi Minh City, Vietnam,
[3]Renesas Electronics, Tokyo, Japan

Today's car information systems (CIS) are growing into integrated cockpit systems, supporting not solely infotainment, such as navigation and AV playing/recording, but also driver assistance, such as surround view systems. Also, in-car video transfer via Ethernet is becoming widespread. Such networks connect camera modules, head unit controllers and rear-seat display units, and carry video signals encoded in H.264 according to EthernetAVB. Thus, it is necessary for integrated cockpit systems to handle significant amounts of video processing. A key requirement for such systems is also low power consumption and thermal management for stable operation.

An SoC for integrated cockpit systems is presented, with a video subsystem resolving the following three issues: (1) Full-HD 12-channel video processing is needed at a maximum, for the following scenario: Blu-ray disc playback (2ch) + rear seat entertainment (4ch) + drive recoding (2ch) + surround view operation (4ch). The video processing should operate in parallel with navigation or/and cognitive processing. Also, the processing relating to driver assistance operating on the function safety (FuSa) OS must not be disturbed by other processing for infotainment on the consumer non-FuSa OS. (2) We must reduce memory bandwidth to support Full-HD 12-channel video processing, which requires more than 20GB/s bandwidth in the worst case. Without this reduction, it would cause performance degradation of navigation/cognitive processing, and an increase in total power consumption. (3) We must achieve low-latency for the surround view system at less than 100ms end-to-end, which is identical to 30cm migration length at 10km/h, so that a driver can control his/her car safely during parking.

Figure 4.4.1 shows the video performance and specifications of the SoC. It achieves 750Mpixels/s, Full-HD 12-channel video codec capability, and supports 4K resolution at H.264 and H.265/HEVC decoding. Moreover, it integrates 3D I/P (interlace-to-progressive) conversion, scaling, blending, color management, skew compensation, and other general video-processing functions, which can process Full-HD 12-channel video. The performance exceeds previous works [1-6], as well as exceeds typical performance improvement trends for application processors (lower left of Fig. 4.4.1). Also, the power consumption at Full-HD H.264 12-channel decoding is measured to be 197mW. The power efficiency, which ranges from 0.16-to-0.29nJ/pixel, is superior to prior works.

Figure 4.4.2 shows a block diagram of the SoC. To achieve Full-HD 12-channel video performance, we implemented 6 types of 17 video processors, indicated by diagonal mesh elements in Fig. 4.4.2, which are connected with each other via hierarchical buses. These processors can perform video processing in real-time and in parallel with navigation or/and cognitive processing in the CPU/GPU. On the other hand, if all processors were to operate at the same time, the peak power consumption would be large, even when the required performance is low. To mitigate this, each processor can adjust its own clock frequency by masking a variable number of cycles periodically. This averages power consumption and reduces the peak when maximum performance is not needed.

Figure 4.4.3 shows the architecture of the video subsystem with memory bandwidth compression. Lossless compression (LLC) does not degrade image quality, but in general, its complicated logic needs more silicon area than that of lossy compression (LC) with small degradation of image quality. In our design, LLC logic consumes 4.5× more area than LC. We observed that a small degradation in image quality is unnoticeable in small car displays. Thus, we implemented LLC and lossless decompression (LLD) in the codec processors (CP) to avoid accumulative noise derived from its processing characteristics, and LC in video processors (VP), blending processors (BP) and display processors (DP) to save silicon area. As a bridge between the lossless and lossy domain, VP can read lossless-compressed pictures. Also, the CPU and GPU have to read lossy-compressed pictures for overlay graphics and other general purposes. Therefore, we located lossy-decompression (LD) logic into memory controllers, enabled by accessing a specific address space. Thus, all modules in the SoC can access both lossy-compressed and non-compressed pictures by switching accessing address. This architecture reduces memory bandwidth from 18.6GB/s to 9.2GB/s in typical Full-HD 12-channel operation. Moreover, the SoC integrates a blocking mechanism so that non-FuSa applications for infotainment never disturb FuSa applications for driver assistance. Although both non-FuSa and FuSa applications can activate the same video processor, FuSa applications can block the register access from non-FuSa applications by setting a specific register available to the FuSa OS only.

Figure 4.4.4 details the lossless compression technology implemented in the SoC. The algorithm is a combination of multi-mode 2D-DPCM and variable-length coding, and its execution unit is 256B, 64pixels×4lines. It compresses 256B of write data to 64×(n+1)B, where n = 0-3, before storing to DDR. Conversely, the read data from DDR is decompressed to 256B and stored to a cache. Although the size of a single-read access with video-codec processing is usually less than 256B, the cache can improve the ratio of valid data in the total data read from DDR, which in turn decreases the number of read accesses. If we adopted a smaller unit of compression to remove the cache, it would degrade DDR access efficiency, because of the overheads associated with accessing DDR for small data sizes. The graph in Fig. 4.4.4 shows that the best unit size is 256B for LPDDR4. Any smaller size would degrade the compression ratio due to reduced DDR access efficiency. However, exceeding 256B also degrades efficiency due to the increase in accesses upon a cache miss. As a result, this scheme, 256B unit compression with the cache, achieves a 70% reduction of memory bandwidth.

Figure 4.4.5 illustrates the approach to low-latency video decoding. Stream processors (SP) and CPs execute video decoding cooperatively. In normal mode for variable-bitrate-encoded streams used for infotainment, the processing time of the SP varies depending on the bitrate of the input stream. To flatten the variance, the SP and CP work independently in time, and the working data between the SP and CP processing is stored in DDR, resulting in several frames latency. A low-latency mode for constant-bitrate-encoded streams is used for video transfer from camera modules to head unit controllers in surround view systems. In this mode, a SP and CP pair work synchronously by sharing a specific FIFO, achieving less than 1ms latency for H.264 decoding. Moreover, the CP can output an IRQ every 16n lines of completed decoding process (n an arbitrary integer). Thereby, the application on the CPU can know which area of a picture is prepared in the DDR and can order rendering processors (RPs) to start their skew-compensation processing in advance of the end of the whole picture-decoding process. This scheme achieves 70ms latency for decoding and skew compensation processing, which is superior to a conventional work [6].

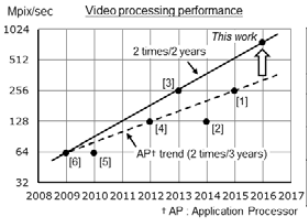
Figure 4.4.6 shows measured results. Since three sets of SPs and CPs are implemented in the SoC, 1 frame of a Full-HD picture must be decoded in 8.3ms for 12-channel processing, and all measured processing time takes less than 8.3ms. Also, the measured power efficiency of decoding H.264 and H.265/HEVC streams ranges from 0.16-to-.29nJ/pixel. The measured power consumption of Full-HD H.264 12-channel decoding is 197mW, which is reduced by 20% with memory bandwidth compression. Fig. 4.4.7 is a micrograph of the video subsystem in the SoC.

References:

[1] C.-C. Ju et al., "A 0.5nJ/Pixel 4K H.265/HEVC Codec LSI for Multi-Format Smartphone Applications," *ISSCC Dig. Tech. Papers*, pp. 336-337, 2015.
[2] A. Wang et al., "Heterogeneous Multi-Processing Quad-Core CPU and Dual-GPU Design for Optimal Performance, Power, and Thermal Tradeoffs in a 28nm Mobile Application Processor," *ISSCC Dig. Tech. Papers*, pp. 180-181, 2014.
[3] C.-T. Huang et al., "A 249Mpixel/s HEVC Video-Decoder Chip for Quad Full HD Applications," *ISSCC Dig. Tech. Papers*, pp. 162-163, 2013.
[4] M. Mehendale et al., "A True Multistandard, Programmable, Low-Power, Full HD Video-Codec Engine for Smartphone SoC," *ISSCC Dig. Tech. Papers*, pp. 226-227, 2012.
[5] Y. Kikuchi, "A 222mW H.264 Full-HD Decoding Application Processor with x512b Stacked DRAM in 40nm," *ISSCC Dig. Tech Papers*, pp. 326-327, 2010.
[6] K. Iwata et al., "A 342mW Mobile Application Processor with Full-HD Multi-Standard Video Codec," *ISSCC Dig. Tech. Papers*, pp. 158-159, 2009.

 978-1-4673-9467-3/16/$31.00

Technology		16nm CMOS
Chip size		16.4mm² (Video)
Supply voltage		0.8V (core)
Clock frequency		400MHz (Video)
External memory		LPDDR4-3200
Video codec	Performance	750Mpixels/sec, 1920x1080 x 30fps x 12channels @H.264
	Maximum resolution	4096x2304 @H.265/HEVC, H.264
	Standard	H.265/HEVC, H.264, MPEG-2/4, VC-1, VP8
Video processing		3D I/P conversion, Scaling, Blending, Color management, Skew compensation, others
Memory bandwidth compression		Lossy (-50% fixed ratio), Lossless (-70% on average)
Power consumption		197mW at Full HD H.264 12ch decoding



Comparison with other works

	This work	ISSCC 2015 [1]	ISSCC 2013 [3]	ISSCC 2012 [4]	ISSCC 2009 [6]
Chip type	AP	AP	Decoder	AP	AP
Technology	16n 0.8V	28n 0.9V	40n 0.9V	32n 1.1V	65n 1.2V
Performance	750 Mpix/s	249 Mpix/s	249 Mpix/s	124 Mpix/s	62 Mpix/s
Core power	0.16-0.29 nJ/pix	0.51 nJ/pix	0.31 nJ/pix	0.52 nJ/pix	5.50 nJ/pix

Figure 4.4.1: Video performance and specifications.

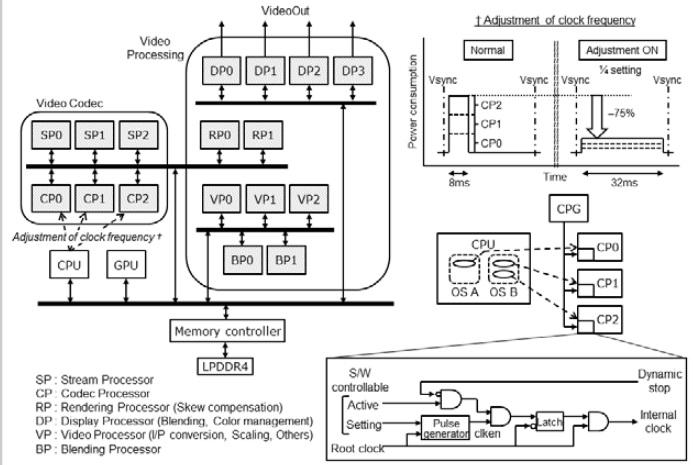


Figure 4.4.2: Block diagram.

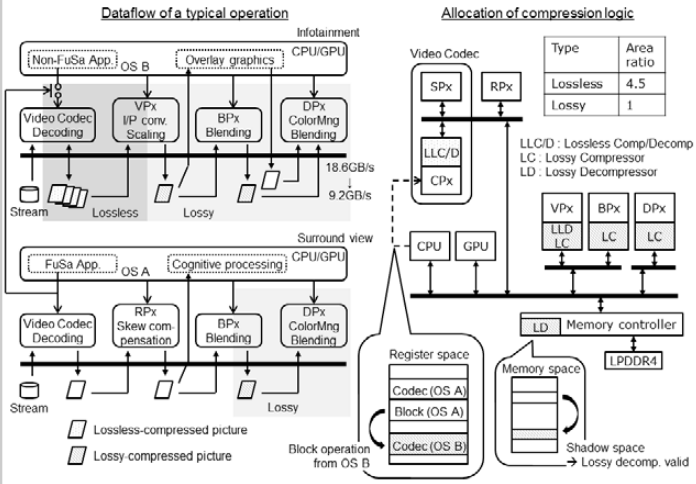


Figure 4.4.3: System architecture.

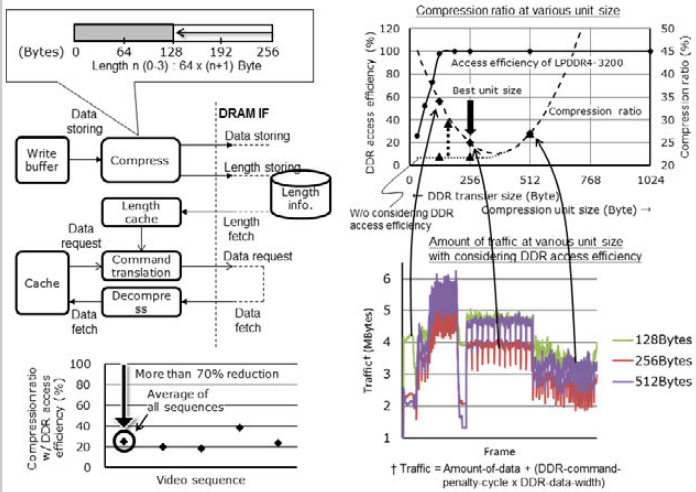


Figure 4.4.4: Memory bandwidth compression.

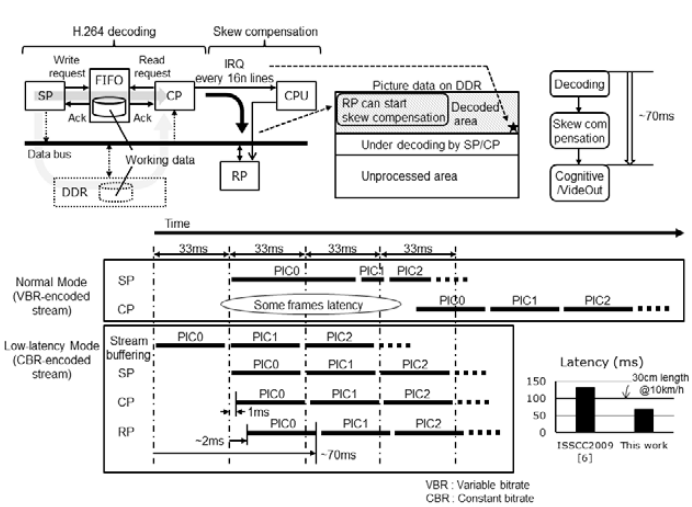


Figure 4.4.5: Low-latency video decoding.

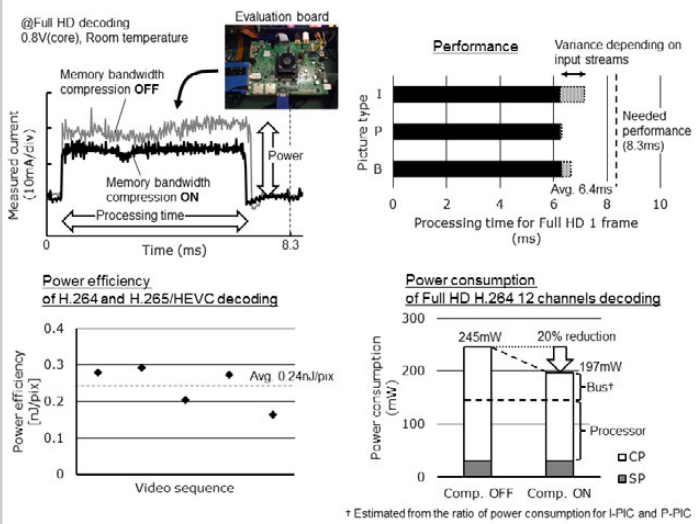


Figure 4.4.6: Measured results.

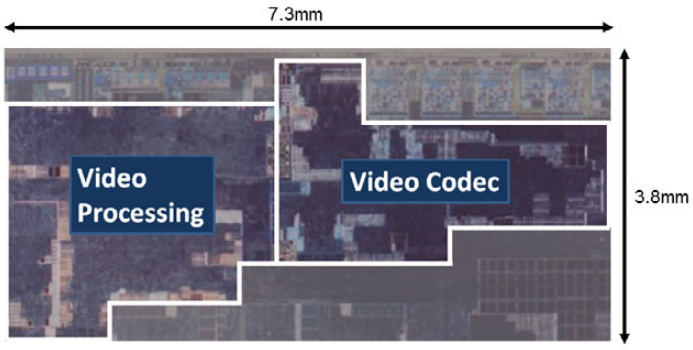


Figure 4.4.7: Chip micrograph.

978-1-4673-9467-3/16/$31.00

4.5 A 16nm FinFET Heterogeneous Nona-Core SoC Complying with ISO26262 ASIL-B: Achieving 10^{-7} Random Hardware Failures per Hour Reliability

Chikafumi Takahashi[1], Shinichi Shibahara[1], Kazuki Fukuoka[1], Jun Matsushima[1], Yuko Kitaji[1], Yasuhisa Shimazaki[2], Hirotaka Hara[2], Takahiro Irita[2]

[1]Renesas System Design, Tokyo, Japan,
[2]Renesas Electronics, Tokyo, Japan

The role of car information systems (commonly referred to as car infotainment) is expanding from dedicated navigation systems to joint car-cockpit systems, including the dashboard meter, telematics for the internet/cloud, and advanced driver-assistance systems (ADASs), such as adaptive cruise control and a pre-crash safety system. The expanding role for car information systems requires higher computational performance, but also safety mechanisms which prevent serious accidents. This paper presents an SoC for the next generation of car infotainment, achieving high performance powered by nine heterogeneous CPUs and a high level of safety, complying with ISO26262 ASIL-B. It has two key features: 1) Run-time test for functional safety, which can detect wear-out faults, such as random fault, time-dependent dielectric breakdown, and electromigration; 2) A killer-droop (critical voltage droop) monitor with droop prediction, which can avoid a delay fault caused by voltage droop.

Figure 4.5.1 shows a block diagram of the 111.36mm^2 chip, including two clusters of quad-core application CPUs, one 3D GPU, one real-time CPU, and built-in self-test (BIST) modules. These BIST modules are the key components of run-time test for functional safety. Functional safety is an approach to reduce the risk of harm to people and property. ISO26262 [1] is an international standard of functional safety for road vehicle electronics systems. It describes a requirement called the automotive safety integrity level (ASIL) and its safety assessment. ASIL is classified from ASIL-A to ASIL-D, based on the safety requirement. One of the important targets is guaranteeing the failure rate measured by failure-in-time (FIT). For example, ASIL-B requires 100 FIT: 1.0×10^{-7} random hardware failures per hour. To decrease the failure rates, introducing "safety mechanisms" are required. Safety mechanisms are defined as technical solutions to detect faults or control failures in order to achieve or maintain a safe state, according to ISO26262. It also defines fault as an abnormal condition that can cause failure. Therefore, detecting faults and returning to a safe state within the fault-tolerant time window (window of time in which failures are caused by faults) are key features of a safety mechanism.

Typically, redundant hardware, self-test by software, and self-test supported by hardware are incorporated into a safety mechanism. Redundant hardware is the most reliable method because it can detect faults by detecting differences between the target and duplicated hardware at the expense of extra hardware. Self-test by software does not need the extra hardware, but it requires a long test time, which causes an interruption in system function at run-time, i.e. run-time self-test (power-on self-test is another self-test method). Hardware support of self-test can reduce test time, but requires some footprint on the chip. In our chip, we use several safety mechanisms to meet application requirements. One CPU in real-time domain applies dual-core lock step (DCLS). DCLS features two identical CPUs, enabling detection of differences in output between the two CPUs. Only DCLS satisfies the interrupt response time required for real-time applications, such as communication for critical situations. CPUs in an application-domain apply the self-test supported by BIST hardware with test time slicing.

Figure 4.5.2 shows the self-test method. The key feature is a flexible test sequence by implementing master BIST controllers for two CPU clusters and the GPU. BIST controllers are implemented for each CPU and GPU domain. Those BIST components utilize the BIST logic for manufacturing test. During the power-on self-test, all CPU clusters and CPU cores can be tested in parallel, which shortens the boot up time. On the other hand, CPU cores can also be tested sequentially and individually during run-time to shorten the blackout time of interrupts. While a CPU is under self-test, it cannot receive any interrupt requests. In addition, our implementation can slice the test suite into multiple pieces so that a maximum interrupt response time can be guaranteed. In this chip, the self-test time without slicing for the largest CPU core is 4.63ms, which exceeds 2.0ms as required in audio applications. Our DCLS implementation consumes over 2.30× the area of a normal single core. On the other hand, our hardware-supported self-test consumes less than 1.1× of the area. If the target application is tolerant toward interrupt blackout time to some extent, this self-test method is advantageous due to its reduced area cost.

Figure 4.5.3 shows the run-time self-test sequence and test time. ISO26262 defines a diagnosis test interval (DTI) which depends on target application. Run-time self-test is executed periodically and it must be finished within the DTI timeframe. At first, the BIST timer sends an interrupt request to the first CPU core. Then, CPUs, the L2 cache (L2$) and snoop controller (SCU) are tested. While a CPU core is being tested, the others can be run normally. And, while L2$+SCU are under test, all CPUs go into sleep mode until the test is completed. The L2$ and SCU feature a RAM initialization function, triggered by reset negation caused by power on and resuming from self-test. We also implement a mask function for RAM initialization and incorporate ECC into the memories, using redundant data for error detection and correction, to exclude them from the periodic self-tests. It can minimize the blackout time for interrupt response. This mechanism can achieve 10^{-7} random hardware failures per hour reliability without exceeding the 2.0ms test time required in audio applications.

This chip has a fault prediction mechanism named a 'killer-droop' monitor. The killer-droop is a critical voltage drop causing a delay fault. Figure 4.5.4 shows the block diagram and its function. This monitor has three key features compared with prior works [2-4]: 1) Voltage sampling at 2.0GHz, which is same as the fastest CPU clock; 2) Voltage-droop prediction; 3) Ability to stop the clock and resume from a low frequency. This monitor measures the differential voltage between core voltage (V_{DD}) and ground (V_{SS}) in the CPUs and outputs it as Vcode (coded differential voltage), which is used in the droop prediction logic. Minimizing the parasitic RC influence in the time-to-digital converter (TDC) by noise isolation and equal-length wiring achieves 5mV voltage resolution. If the predicted Vcode is lower than a threshold Vcode, corresponding to the minimum required to prevent a delay fault, the clock gating cell stops the clock supply to the target CPUs. After the supply voltage becomes stable, as the result of the power reduction associated with the clock gating, the CPU clock controller resumes the clock supply with 1/32 the frequency of the original clock and increases the clock frequency gradually. Resuming clock supply with full speed causes excessive voltage droop. When the threshold Vcode (killer-droop threshold) is small, the voltage variation can be minimized. However, the clock supply is stopped frequently, causing performance degradation.

Figure 4.5.5 shows the droop-prediction logic and its impact. This logic accurately predicts Vcode 4-cycles in advance, as shown in the comparison of the predicted Vcode and measured Vcode. The combination of fault prediction and prevention is more effective than fault detection because it can continue the current task without retries. Figure 4.5.6 shows the evaluation results and a comparison to previous works. The X and Y-axis in the graph are the voltage window (difference between maximum and minimum voltage) and the CPU performance normalized by the measured result without clock stoppage. Even when the voltage window is reduced by 10% and 20%, the performance degradation is only 1% and 1.5%, respectively. This result demonstrates killer-droop can be avoided while maintaining performance. Compared with previous works, this monitor defends against killer-droop with no latency.

References:

[1] The International Organization for Standardization, "ISO 26262: 2011 Road vehicles - Functional Safety," 2011.
[2] M. Igarashi et al., "A 28nm HPM Heterogeneous Multi-Core Mobile Application Processor with 2GHz Cores and Low-Power 1GHz Cores," *ISSCC Dig. Tech. Papers*, pp. 178-179, 2014.
[3] A. Grenat et al., "Adaptive Clocking System for Improved Power Efficiency in a 28nm x86-64 Microprocessor," *ISSCC Dig. Tech. Papers*, pp. 106-107, 2014.
[4] K. Bowman et al., "A 16nm Auto-Calibrating Dynamically Adaptive Clock Distribution for Maximizing Supply-Voltage-Droop Tolerance Across a Wide Operating Range," *ISSCC Dig. Tech. Papers*, pp. 152-153, 2015.

 978-1-4673-9467-3/16/$31.00

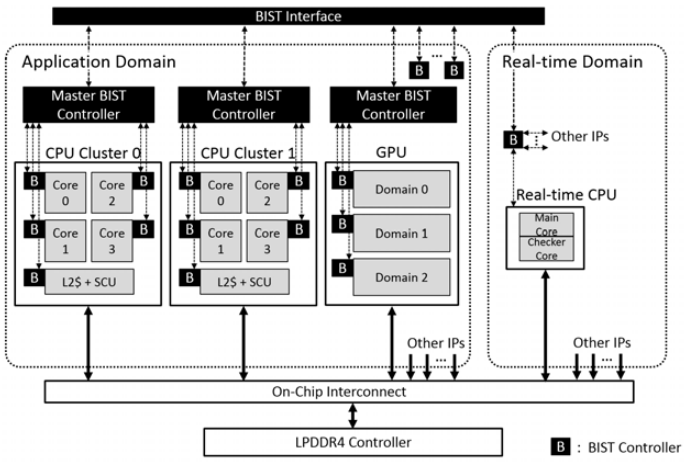
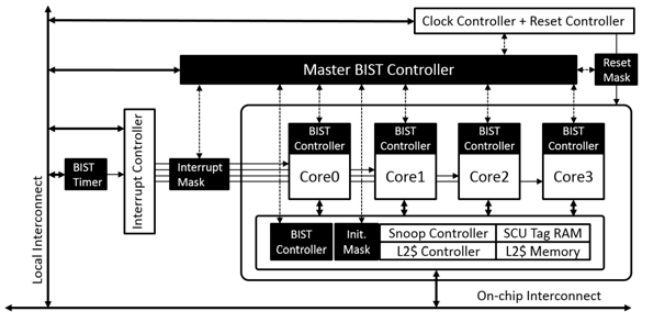


Figure 4.5.1: Block diagram of the SoC.



	No measure	Dual Core Lock Step (Redundant HW)	Software Self Test	Hardware Supported Self-Test	This work
Fault detection coverage	N/A	High	Low	Medium	Medium
Relative logic size of CPU cluster	1.00	>2.30	1.00	<1.10	<1.10
Blackout time of interrupts	N/A	N/A	>100ms	> 10ms	< 2.0ms

Figure 4.5.2: Self-test supported by BIST hardware.

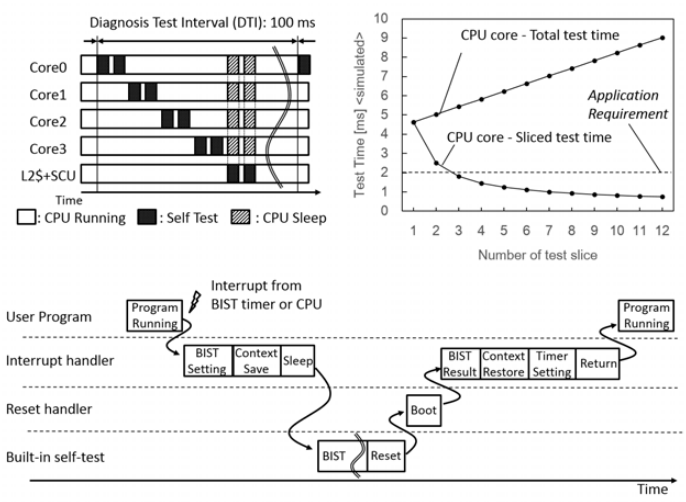


Figure 4.5.3: Test time slicing.

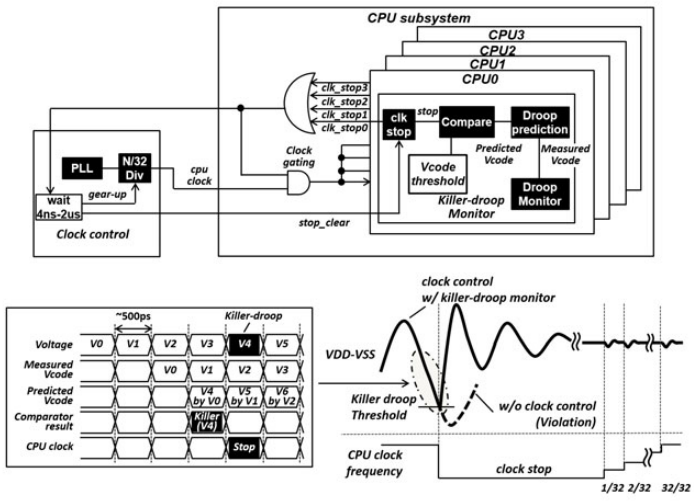


Figure 4.5.4: Adaptive clock control for delay fault.

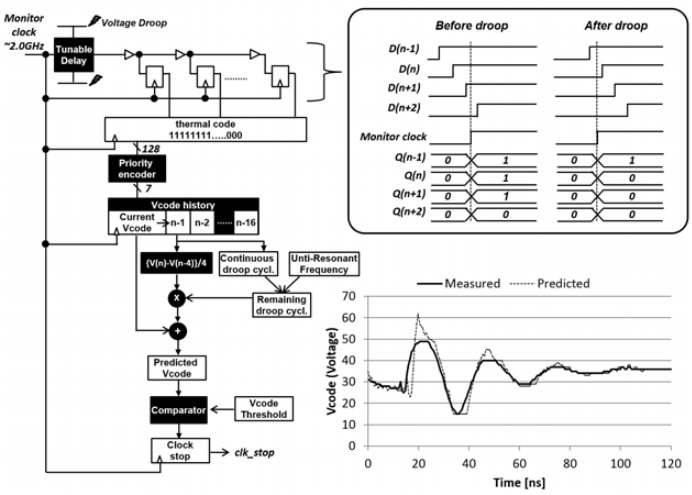
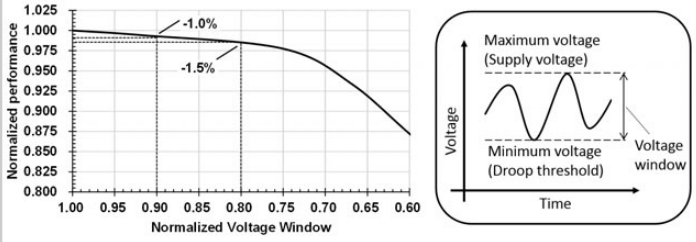
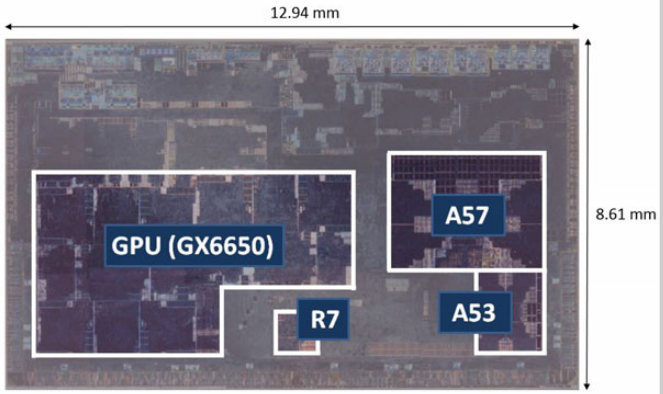


Figure 4.5.5: Fault prediction with droop monitoring.



	This work	ISSCC2014[2]	ISSCC2014[3]	ISSCC2015[4]
Technology	16nm	28nm	28nm	16nm
Application	Automotive	Mobile	PC	Mobile
Droop Monitor	TDC (Time to Digital converter)	Ring oscillator	DLL	Replica path delay
Monitor Interval	1-cycle	50ns	1-cycle	1-cycle
Reaction Latency	0-cycle	100ns	2-cycle	1.5-cycle
Reaction Mechanism	Clock stop	Frequency down	Clock pulse stretch	Inserting delay to clock tree
Droop Prediction	Yes	n/a	n/a	n/a

Figure 4.5.6: Performance impact and comparison.



Process	TSMC 16nm FinFET
Size	12.94 mm x 8.61mm, 111.36mm^2
Memory	LPDDR4-3200

CPU	Quad Cortex-A57 2.0GHz
	Quad Cortex-A53 1.2GHz
	Dual-core Lock-step Cortex-R7 800MHz
GPU	PowerVR GX6650

Figure 4.5.7: Die photo and specifications.

978-1-4673-9467-3/16/$31.00

4.6 A 65nm CMOS 6.4-to-29.2pJ/FLOP@0.8V Shared Logarithmic Floating Point Unit for Acceleration of Nonlinear Function Kernels in a Tightly Coupled Processor Cluster

Michael Gautschi[1], Michael Schaffner[1], Frank K. Gürkaynak[1], Luca Benini[1,2]

[1]ETH Zurich, Zurich, Switzerland, [2]University of Bologna, Bologna, Italy

Energy-efficient computing and ultra-low-power operation are requirements for many application areas, such as IoT and wearables. While for some applications, integer and fixed-point processor instructions suffice, others (e.g. simultaneous localization and mapping – SLAM, stereo vision, nonlinear regression and classification) require a larger dynamic range, typically obtained using single/double-precision floating point (FP) instructions. Logarithmic number systems (LNS) have been proposed [1,2] as an energy-efficient alternative to conventional FP, as several complex operations such as MUL, DIV, and EXP translate into simpler arithmetic operations in the logarithmic space and can be efficiently calculated using integer arithmetic units. However, ADD and SUB become nonlinear and have to be approximated by look-up tables (LUTs) and interpolation, which is typically implemented in a dedicated LNS unit (LNU) [1,2]. The area of LNUs grows exponentially with the desired precision, and an LNU with accuracy comparable to IEEE single-precision format is larger than a traditional floating-point unit (FPU). However, we show that in multi-core systems optimized for ultra-low-power operation such as the PULP system [3], one LNU can be efficiently shared in a cluster as indicated in Fig. 4.6.1. This arrangement not only reduces the per-core area overhead, but more importantly, allows several costly operations such as FP MUL/DIV to be processed without contention within the integer cores without additional overhead. We show that for typical nonlinear processing tasks, our LNU design can be up to 4.2× more energy efficient than a private-FP design.

For an accurate comparison, we have manufactured and measured two separate chips with a quad-core cluster system using the UMC65nm LL technology. Each chip contains an identical cluster with four 32b OpenRISC cores, a 16kB shared tightly coupled data memory (TCDM) and 1kB instruction caches. In the 1st chip, one LNU using an 8b integer, 23b fractional and a sign bit LNS format is shared by four cores, and in the 2nd chip, all four cores have their own private FPU. A third shared FPU design has also been evaluated, but not included in this comparison as it was much slower (up to 46%) due to contention (up to 96%). LNU and FPUs have been tightly integrated into the integer ALUs of the processors, as illustrated in Fig. 4.6.1 and can be accessed using standard OpenRISC FP instructions. The FPU used for comparisons in this work is IEEE single-precision compliant and supports FP ADD/SUB/MULT and casts and is based on a multiply-add architecture with a shared normalizer. The measured FPU energy efficiency @0.8V of 15.3-to-18.4pJ/op is compatible with other state-of-the art designs [5,6].

The main challenge of the LNU circuit shown in Fig. 4.6.2 is to efficiently implement the two nonlinear functions (f+ and f- as shown in Fig. 4.6.3) that are needed to calculate ADD/SUB in LNS. These two functions are approximated using a combination of LUTs and interpolators. To achieve IEEE single-precision accuracy with reasonable area overhead, two separate blocks are used. The main 2^{nd}-order polynomial interpolator handles the regions where f+ and f- are almost linear and the corresponding LUTs have been partitioned into logarithmically spaced segments, using only 1k and 2.1k entries for f+ and f-, respectively. Each set of coefficients has been optimized using a finite-precision-aware minimax fitting algorithm to minimize the bit widths (from 3b to 32b) and the amount of required coefficient samples at the given precision requirement. For results that fall in the critical region (Fig. 4.6.3), a novel co-transformation block has been designed, which employs a mathematical transformation to circumvent precision problems [1,2]. First it evaluates $(1-2^z)/(1+2^z)$ which can be efficiently approximated using a first-order polynomial. The result is then fed to a log2 block, which has been implemented with the HOTBM method from [4]. In order to minimize LUT size, the log2 domain has been reduced to [1,2), and the output of $(1-2^z)/(1+2^z)$ is calculated in shifted format by storing the 0^{th}-order coefficients in pre-shifted format. The coefficients are then properly aligned and processed. This arrangement allows the employed zero counter, shifter and log2 block to be used to calculate typecasts and native log2 functions as well. The pre-processing block of the LNU decodes the command, chooses the appropriate interpolator block and gates the input operators. Evaluations showed that input silencing is critical for low-energy operation and reduces energy consumption for LNU additions from 107.7pJ/op to only 28.7pJ/op in our design. Finally, the post-processing block combines all intermediate results and performs special case handling such as over-/underflows. Our LNU implements more functions (casts, exp, log), uses far smaller LUTs (14.1kB instead of 22.9kB) and reduces the area overhead by 35% when compared to the most advanced design in the literature [2].

Both units are tightly integrated in the processor's data path as shown in Fig. 4.6.1 and share a write-back port of the register file with the load-store unit. The FPU requires only one pipeline stage, while the LNU requires three stages. The shared LNU is managed by a fair round-robin arbiter. The shared LNU is only needed to process LNS ADD, SUB, LOG, EXP and casts. Many LNS operations such as MUL, DIV, SQRT and comparisons can be directly computed in the integer ALU of the cores in a single cycle, which is energy efficient, reduces LNU contention, and makes it more attractive for sharing in a multi-core setting. The efficiency of several LNS and FP-instructions is compared in Fig. 4.6.4. While LNS ADD/SUB are less energy efficient than the FP equivalents, the LNS MUL requires 36% less energy than in FP. Apart from these basic instructions, LNS supports extremely energy efficient, single cycle square-roots (6.4pJ/op) and divisions (12.1pJ/op) utilizing the shifter and adder of the ALU with dedicated special case handling. Also, complex functions such as 2^x and $\log_2(x)$ can be computed in the LNU in four cycles for 13.3 and 22.6pJ/op, respectively.

Both architectures have been benchmarked using a suite consisting of linear algebra kernels, matrix decompositions and more complex, nonlinear functions involving projective transforms, radial basis functions, trigonometric functions and distance computations. The benchmark set has been generated using MATLAB and its C++ embedded coder to ensure that competitive kernel implementations are used. The LLVM compiler has been adapted to support the LNS format for OpenRISC, letting the compiler handle automatically low-level details, such as LNU latency. The upper part of Fig. 4.6.5 shows kernel characteristics such as FP instruction ratios, codesize and IPC where the lower part shows kernel execution time with its power consumption and energy savings. For complex kernels, such as 3D-distance computations and Cholesky decompositions, the LNU can make use of its extended ISA (DIV, SQRT) allowing such applications to run 4.2× more energy efficiently, as illustrated in Fig. 4.6.6. A big difference is observed when EXP and LOG functions are frequently used because they can be performed natively on the LNU, while the FPU uses software emulation consuming 51 (EXP) and 85 (LOG) cycles.

For pure linear-algebra kernels, AXPY, GEMM, and GEMV, the cluster with private FPUs is 5-to-13% more energy efficient than the shared LNU due to LNU's longer latency for FP add/sub. This relatively small gap even for ADD-SUB intensive benchmarks is easily amortized for complex algorithms with many multiplications, divisions and nonlinear functions. The utilization of the shared LNU on our benchmarks was 0.37, on average, with a maximum of 0.61 where the high utilization led to 10% stalls due to access contention. On average, such contention only occurred in 4% of all FP operations. The chips with private FPU and shared LNU integrated in a multi-core cluster are shown in Fig. 4.6.7. The two chips are comparable in size, but the shared LNU, with a top energy efficiency of 6.4pJ/FLOP @0.8V, permits computation of complex nonlinear kernels up to 4.2× more energy efficiently than the FPU cluster.

Acknowledgments:
We thank M. Burger, T. Gautschi, L. Müller, Y. Popoff, F. Scheidegger, and F. Schuiki for their valuable work and commitment during their semester projects. This research was supported by the IcySoC project, evaluated by the Swiss NSF and funded by Nano-Tera.ch with Swiss Confederation financing.

References:
[1] J. Coleman et al., "The European Logarithmic Microprocessor," *IEEE Trans. on Computers*, vol. 57, no. 4, pp. 532-546, April 2008.
[2] R.C. Ismail et al., "ROM-less LNS," *IEEE Symp. on Computer Arithmetic*, pp. 43-51, 2011.
[3] D. Rossi et al., "A -1.8V to 0.9V Body Bias, 60 GOPS/W 4-core Cluster in Low-Power 28nm UTBB FD-SOI Technology," *IEEE SOI-3D-Subthreshold Microelectronics Tech. United. Conf.*, 2015.
[4] J. Detrey et al., "Table-Based Polynomials for Fast Hardware Function Evaluation," *IEEE Int'l Conf. on Application-Specific Systems, Architectures and Processors*, pp. 328-333, 2005.
[5] S. Galal et al., "Energy-Efficient Floating-Point Unit Design," *IEEE Trans. On Computers*, vol. 60, no. 7, pp. 913-922, July 2011.
[6] H. Kaul et al., "A 1.45 GHz 52-to-162GFLOPS/W Variable-Precision Floating-Point Fused Multiply-Add Unit with Certainty Tracking in 32nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 182-184, 2012.

 978-1-4673-9467-3/16/$31.00

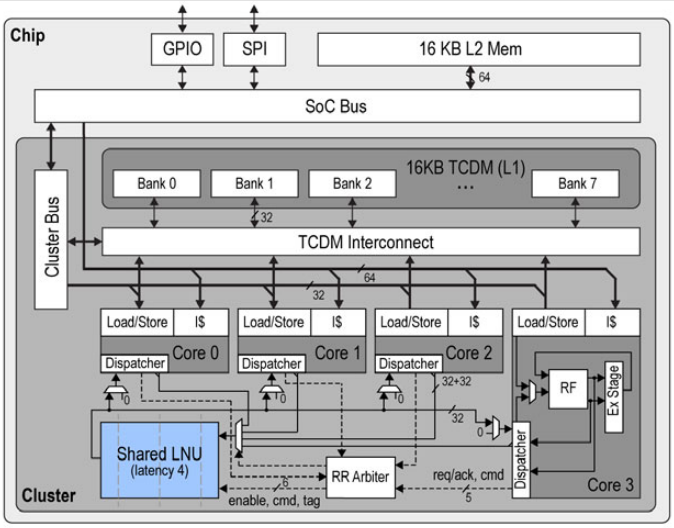


Figure 4.6.1: Four-core cluster architecture with one shared LNU directly integrated in the pipeline of the processors.

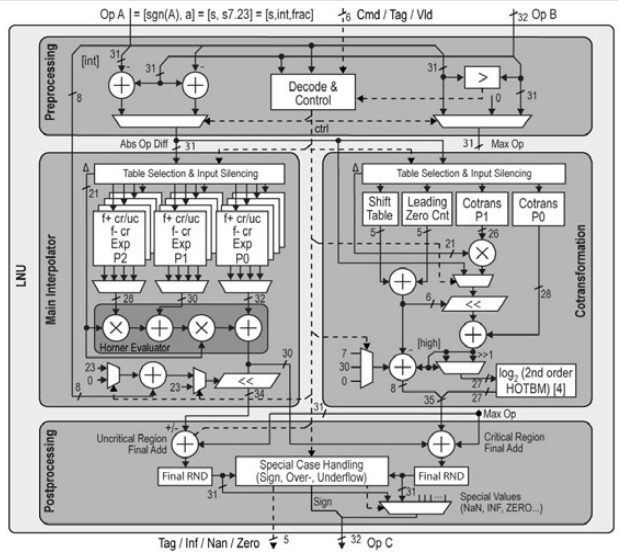
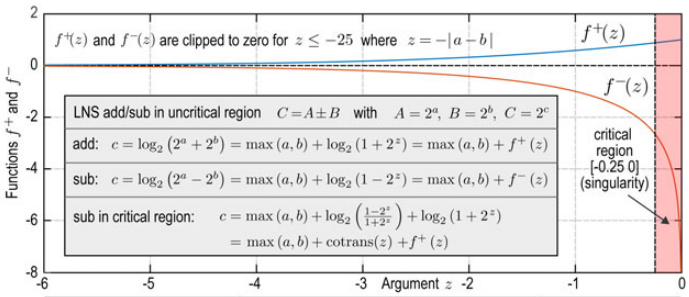


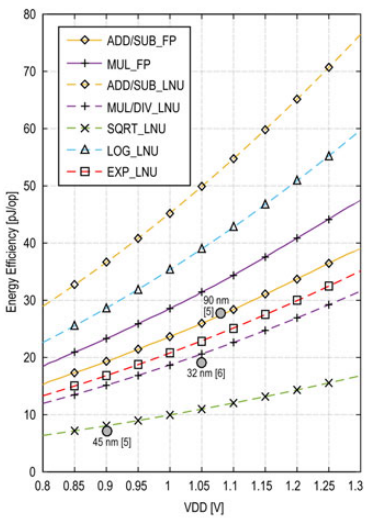
Figure 4.6.2: LNU architecture with main interpolator and co-transformation.



LNU Comparison	LNU	ELM [1,2]	ROM-less [2]
functionality	ftoi, itof, add, sub, exp, log	add, sub	add, sub
precision (max/avg err) [ulp] [1]	0.479 / 0.177	0.454 / 0.178	0.498 / 0.174
LUT size [entries] / [Kbit]	5.1k / 113.1	8.2k / 256.4	4k / 183.3
technology	65 nm	180 nm	180 nm
area [squm]	57'264	904'943	589'357
Total area [kGE]	62	-	-
Post-synthesis (w/o registers) [kGE]	40	97	63
min delay [ns]	6	11.74	7.10
max delay [ns] (critical region)	6	13.15	14.79

[1] precision compared to standard floating point as calculated in [3]

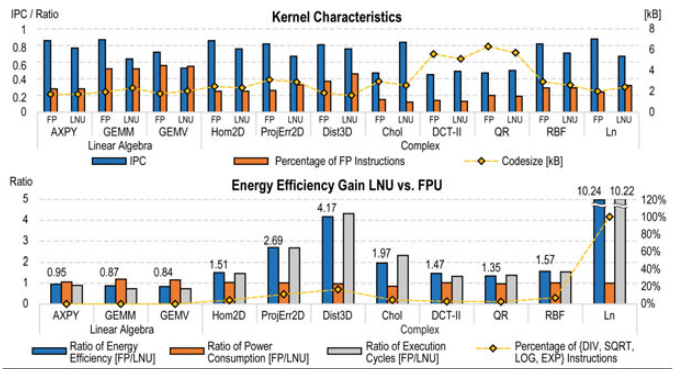
Figure 4.6.3: LNS f^+ / f^- functions with singularity in the critical region where the cotransformation applies. Comparison with related work.



Energy / Instruction	FPU [1]	LNU [1]	Horowitz [5]		Intel [6]
Technology	65nm LVT	65nm LVT	90nm	45nm	32nm
Vdd [V]	0.8	0.8	1.08	0.9	1.05
Pipeline Depth	2	4	10	6	3
Area [mm²]	0.015	0.089	0.113	0.016	0.045
ADD [pJ/FLOP]	15.3	28.7	-	-	-
SUB [pJ/FLOP]	15.2	29.2	-	-	-
MUL [pJ/FLOP]	18.4	11.8	-	-	-
DIV [pJ/FLOP]	-	12.1	-	-	-
Sqrt [pJ/FLOP]	-	6.4	-	-	-
EXP [pJ/FLOP]	-	13.3	-	-	-
LOG [pJ/FLOP]	-	22.6	-	-	-
Fused MUL/ADD (FMA) [pJ/FLOP]	-	-	27.7 [2]	7.2 [2]	19.2

[1] background power (NOP) subtracted, includes core activity
[2] FPU datapath only

Figure 4.6.4: Energy efficiency comparison of different FP- and LNU-instructions.



Kernel	Details	Kernel	Details
AXPY	BLAS1 kernel, [100x1] vectors	Chol	Cholesky Decomposition, [10x10] matrices
GEMV	BLAS2 kernel, [10x10] matrices	DCT-II	32-point DCTs
GEMM	BLAS3 kernel, [10x10] matrices	QR	QR Decomposition, [10x10] matrices
Hom2D[1]	Evaluation of 2D homographies (projective transforms)	LN	Element-wise Natural Logarithm
ProjErr2D[1]	Calculation of reprojection error of 2D projective transforms	RBF	Evaluation of a 2D regression function with 25 Gaussian kernels
Dist3D	Distance computations in 3D		

[1] for more details: "Multiple View Geometry", R. Hartley and A. Zisserman, 2000, ISBN 0521540518

Figure 4.6.5: Energy, speedup, and power comparison of shared LNU vs. private FPU.

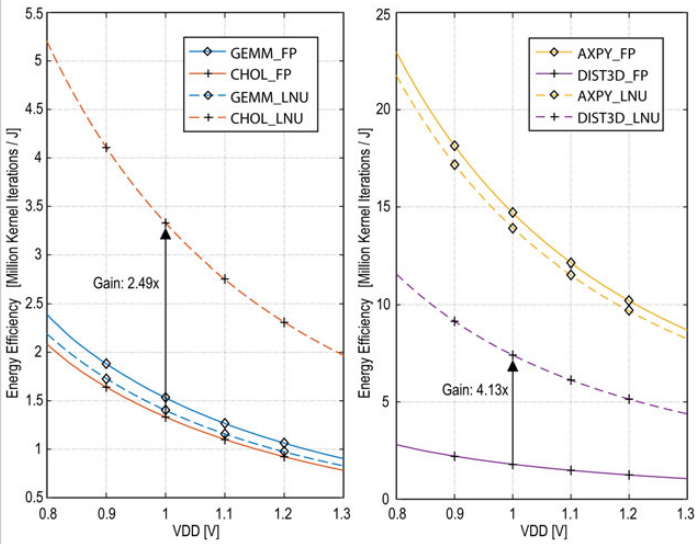
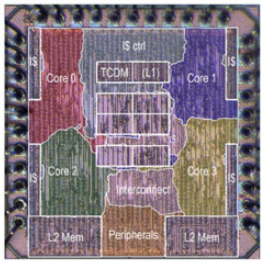
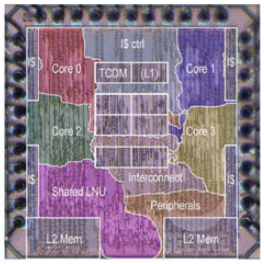


Figure 4.6.6: Energy efficiency in kernel iterations per J at different V_{DD} levels.

Private FPU



Shared LNU



Implementation Details		Private FPU	Shared LNU	ELM [1,2]
Technology		65nm LVT	65nm LVT	180nm
max speed [MHz]		374	337	125
max. Throughput [GFLOPS]		1.1	0.9	0.084
Power @100MHz, 1.2V, 25°C [mW]		41.84	44.0	-
Leakage @ 1.2V, 25°C [mW]		2.823	3.019	-
Best FP-operator efficiency @0.8V [pJ/FLOP]		15.2	**6.4**	-
Precision (max err) [ulp]		0.5	0.478[1]	0.454[1]
avg. lnu/fpu utilization		0.21	0.37	-
Total area [kGE]		719	749	-
Single core area [kGE]		51.1[2]	44.5	-
Instruction support		**Private FPU**	**Shared LNU**	**ELM [1,2]**
Latency add/sub/casts	hw	2/2/2	4/4/4	3/3(4)/ -
Latency mul/div/sqrt [3]	hw	2/-/-	1/1/1	1/1/1
	sw	-/62/56	-/-/-	-/-/-
Latency exp/log [3]	hw	-/-	4/4	- / -
	sw	51/85	-/-	-/-

[1] precision compared to FP [3] [2] including FPU [3] sw emulations: *div* (range reduction, linear initial estimate and 3 Newton iterations), *sqrt* (fast-inverse sqrt and 3 Newton iterations), *log* and *exp* (range reduction, polynomials from "Computer Approximations", J. Hart, 1978, ISBN 0882756427)

Figure 4.6.7: Chip photos and datasheet.

978-1-4673-9467-3/16/$31.00

4.7 A 65nm ReRAM-Enabled Nonvolatile Processor with 6× Reduction in Restore Time and 4× Higher Clock Frequency Using Adaptive Data Retention and Self-Write-Termination Nonvolatile Logic

Yongpan Liu[1], Zhibo Wang[1], Albert Lee[2,3], Fang Su[1], Chieh-Pu Lo[2], Zhe Yuan[1], Chien-Chen Lin[2], Qi Wei[1], Yu Wang[1], Ya-Chin King[2], Chrong-Jung Lin[2], Pedram Khalili[3], Kang-Lung Wang[3], Meng-Fan Chang[2], Huazhong Yang[1]

[1]Tsinghua University, Beijing, China,
[2]National Tsing Hua University, Hsinchu, Taiwan,
[3]University of California, Los Angeles, CA

With the rising importance of energy efficiency, zero leakage power and instant-on capability are highly desired features in energy harvesting sensors, as well as "normally off" high performance processors. However, intermittent power in such systems requires nonvolatile memory (NVM) to hold intermediate data and avoid rollbacks. Previous work has adopted FeRAM and STT-MRAM to achieve zero-standby power and fast-restore nonvolatile processors (NVPs) [1-3]. Previous NVPs, however, suffer from several drawbacks: 1) Various power interrupt periods are not considered; 2) the 2-macro memory architecture slows access speed; 3) worst-case store/restore operations are always performed. We present a 65nm fully-CMOS-logic-compatible ReRAM-based NVP achieving time/space-adaptive data retention. A 1-macro nvSRAM with self-write-termination (SWT) is integrated to boost clock frequency and reduce store energy. The adaptive retention and SWT strategy relieve the ReRAM write endurance challenge (10^6-10^{12}), making it sufficient for most applications. The NVP operates at 100MHz with 20ns/0.45nJ restore time ($T_{RESTORE}$)/energy ($E_{RESTORE}$), realizing 6× reduction in $T_{RESTORE}$, >6000× reduction in $E_{RESTORE}$ and 4× higher clock frequency compared with existing designs.

Figure 4.7.1 shows the challenges of a conventional NVP. When power failures happen, the data in volatile registers and memory is stored into nonvolatile flip-flops (nvFFs) and a NVM macro. However, challenges exist: 1) Previous nonvolatile controllers (NVCs) lack time-domain-adaptive retention for variable power-interrupt periods. For power disruptions longer than a "breakeven time", NVM store should be used for zero leakage, while fast but leaky data retention by lowering V_{DD} is preferred for short power interrupts. Furthermore, faster restore speed is always desired to catch critical events, while store speed can be relaxed by an energy buffer capacitor C_{bulk}; 2) The 2-macro architecture attaches slow NVM on the bus, leading to lower performance in normal mode. In restore operations, data has to be read from NVM to SRAM sequentially, degrading restore speed/energy by 2-to-4 orders of magnitude; 3) Conventional NVPs adopt worst-case data retention, assuming all contents in registers and memory should be updated. However, in real applications, data values in up to 90% NVM bits match previous ones and many bits are in fact unused.

Figure 4.7.2 shows the architecture of the fabricated NVP, consisting of an adaptive NVC, a code ReRAM macro, adaptive nvFFs and a configurable nvSRAM. The NVC workflow is described as follows. The time-domain controller detects the sleep signal, and uses a 2b predictor to determine whether the retention or store operation should be performed. If the store operation is adopted, the NVP backs up volatile data and enters the OFF state. Otherwise, it goes to RETENTION mode. If the retention time exceeds a specific threshold, a timeout signal is generated and the NVP enters STORE mode. When the wakeup signal is active, the NVP goes back to NORMAL mode. During the store/restore operation, the space-domain controller manages the memory size by generating the variable ADDR for nvSRAM.

Figure 4.7.3 presents the adaptive nvFF with SWT. The ReRAM device is switched to a low/high resistance state (LRS/HRS) by applying a SET/RESET voltage (V_{SET}/V_{RESET}) for period T_{SET}/ T_{RESET}. The nvFF has 4 modes: NORMAL, STORE, RESTORE and RETENTION. In NORMAL mode (RSWL=0), it acts as a typical flip-flop. In STORE operation, the data is moved into two ReRAM devices (RL and RR). As ReRAM devices suffer from a wide distribution in T_{SET}/T_{RESET} [5], SWT senses the NX/Q voltage and terminates the STORE operation, suppressing wasted store energy and degraded reliability due to over-RESET/SET for fast-switch cells. In the RESTORE operation, Q becomes logic-0/1 when RL=LRS/HRS. In RETENTION mode, the clock is gated and V_{DD} is lowered to 0.4V to reduce leakage, while most of other domains are power gated. This mode reduces store energy and extends the lifetime of ReRAM devices by avoiding frequently writing to ReRAM in short power interruptions. A scan chain is used to access nvFFs for testability.

Figure 4.7.4 shows the adaptive nvSRAM, which has three modes: SRAM, STORE and RESTORE. In SRAM mode (RSWL=0), the 7T1R-nvSRAM [4] has the same high-speed read/write behavior as a nominal 6T-SRAM. In RESTORE mode, by using a dual-supply-initialization pulse-overwrite (DSI-POW) scheme, Q will become logic-0 (logic-1) if ReRAM is LRS (HRS). The adaptive restore controller trades off restore speed and peak current by configuring restore parallelism (1/4/16WL, i.e. 1×16B, 4×16B, or 16×16B). In STORE mode, nvSRAM cells in the same page store data from SRAM to ReRAM in parallel. However, there would be a large store-DC-current causing nontrivial wasted energy if the store conditions are continuously applied for the worst-case store, determined by the slowest cell. This work incorporates the SWT circuit into each column of nvSRAM array for energy savings. At the beginning of storing, CVDDQ=CVDDQB=V_{SET}-V_{RESET} and BL=BLB=1, leading to SWT being initialized. When WL=1, the BL of a 0-cell (Q=0) drops, causing RSL=V_{SET} for a SET operation. After ReRAM devices finish the SET operation and switch to LRS, the large $I_{STORE\text{-}DC}$ raises the voltage at node Q, and QB flips from 1 to 0. BLB drops and Driver_EN=0, which terminates the SET operation. Similarly, the RESET operation is performed for a 1-cell.

Figure 4.7.5 shows store/restore time/energy of the adaptive NVP. The store energy is reduced via three adaptive strategies: nvSRAM sizing, SWT, and time-domain adaptive retention, which can be applied in a stacking way. The nvSRAM sizing configures the capacity (0/16B/256B/1KB/4KB) according to application requirements. It achieves up to 28× energy savings by eliminating store operations in unused memory. The SWT circuits in the nvFF and nvSRAM avoid store operations in matched ReRAM cells and reduce store energy by up to 172×. A matched cell is defined as a ReRAM device, whose previous value equals the data to be written. The typical ratio of matched cells in NVPs is larger than 80% for embedded benchmarks. Finally, the time-domain adaptive retention saves as much as 2× store energy if most power gating periods are shorter than the breakeven time.

Figure 4.7.6 shows the measurement results. An 8b NVP was implemented using a 65nm CMOS process and logic-process-compatible contact ReRAM [6]. It runs a counter program at 100MHz with a data waveform displayed by GPIOs. The restore time is 170ns under a 4KB nvSRAM configuration, while 20ns is achieved under a 16B nvSRAM configuration. The Shmoo figure of the nvSRAM shows it operates at a wide supply range of 0.4-to-1V. The ReRAM-based NVP is 4× faster than state-of-the-art NVPs benefiting from the 1-macro hybrid memory architecture. The adaptive data retention scheme achieves 6× speedup and >6000× energy reduction in restore operations. Figure 4.7.7 shows the die photograph and summary table.

Acknowledgements:
Supported in part by 863 Project 2013AA01320, YETP0102, Beijing Advanced Innovation Center for Chip Excellence, and NSFC Grant #61271269.

References:
[1] Y. Wang et al., "A 3us Wake-up Time Nonvolatile Processor Based on Ferroelectric Flip-Flops," *European Solid-States Circuits Conf.,* pp. 149-152, 2012.
[2] S. Bartling et al., "An 8MHz 75μA/MHz Zero-Leakage Non-Volatile Logic-Based Cortex-M0 MCU SoC Exhibiting 100% Digital State Retention at VDD=0V with <400ns Wakeup and Sleep Transitions," *ISSCC Dig. Tech. Papers*, pp. 432-433, 2013.
[3] N. Sakimura et al., "A 90nm 20MHz Fully Nonvolatile Microcontroller for Standby- Power-Critical Applications," *ISSCC Dig. Tech. Papers*, pp.184-185, 2014.
[4] A. Lee et al., "RRAM-based 7T1R Nonvolatile SRAM with 2x Reduction in Store Energy and 94x Reduction in Restore Energy for Frequent-Off Instant-On Applications." *IEEE Symp. VLSI Circuits*, pp. C76-C77, 2015.
[5] M.-F.Chang et al., "Embedded 1Mb ReRAM in 28nm CMOS with 0.27-to-1V Read Using Swing-Sample-and-Couple Sense Amplifier and Self-Boost-Write-Termination Scheme." *ISSCC Dig. Tech. Papers*, pp. 332-333, 2014.
[6] M.-F. Chang et al., "A 0.5V 4Mb Logic-Process Compatible Embedded Resistive RAM (ReRAM) in 65nm CMOS Using Low Voltage Current-Mode Sensing Scheme with 45ns Random Read Time," *ISSCC Dig. Tech. Papers*, pp. 434-435, 2012.

 978-1-4673-9467-3/16/$31.00

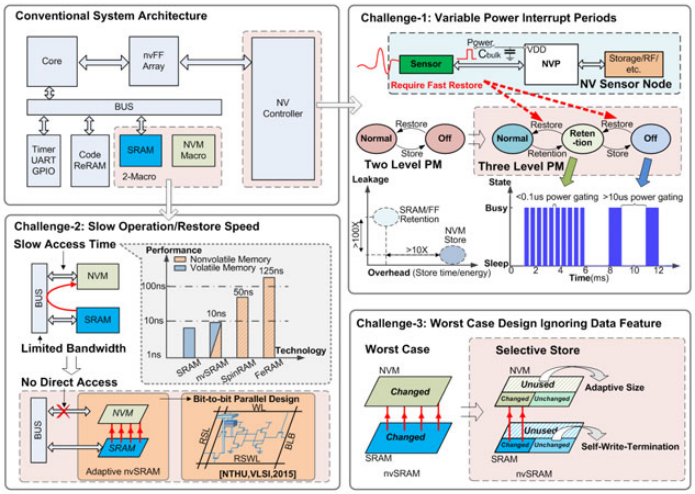


Figure 4.7.1: Challenges of a conventional NVP.

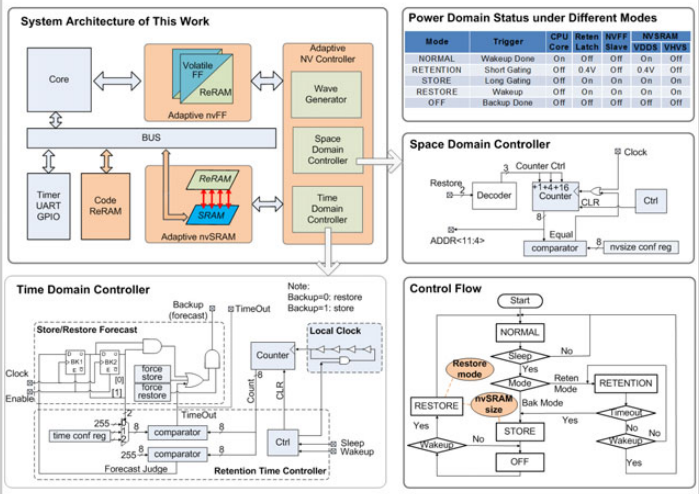


Figure 4.7.2: The fabricated NVP with adaptive data retention.

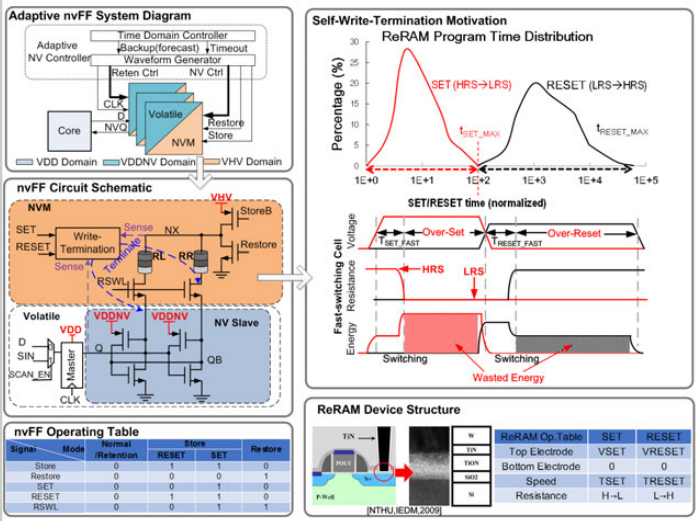


Figure 4.7.3: Adaptive nvFF with data retention and self-write-termination (SWT).

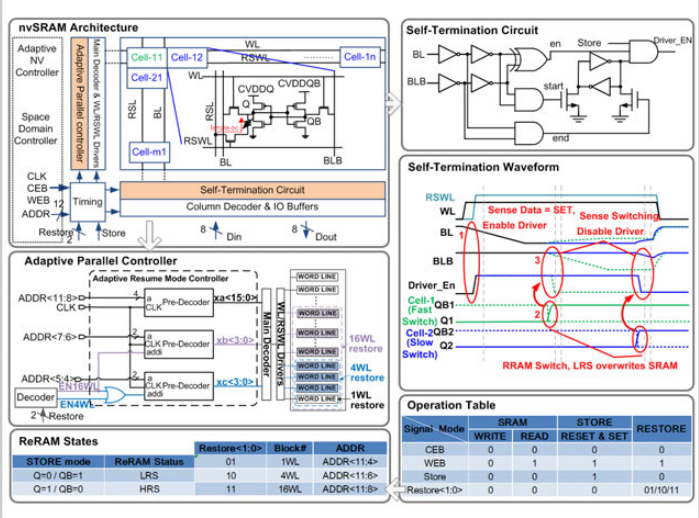


Figure 4.7.4: Adaptive nvSRAM with reconfigurable size and self-write-termination (SWT).

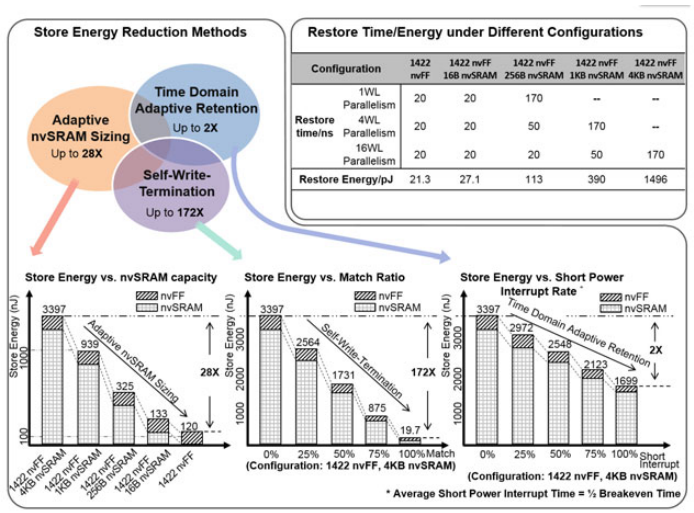


Figure 4.7.5: Restore/store time/energy of the adaptive NVP.

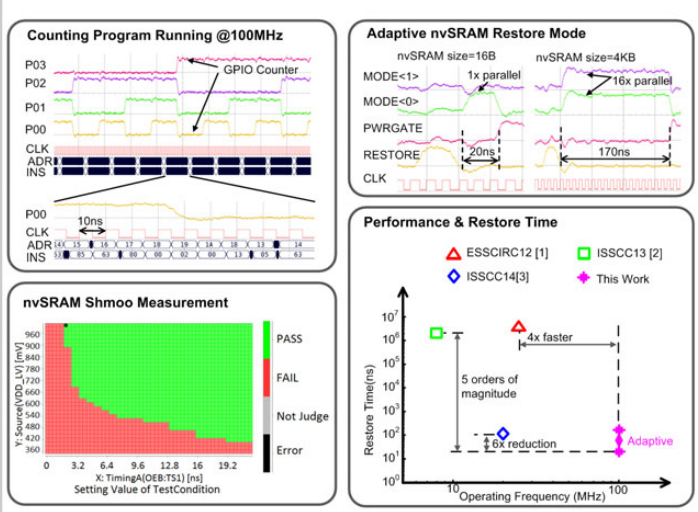
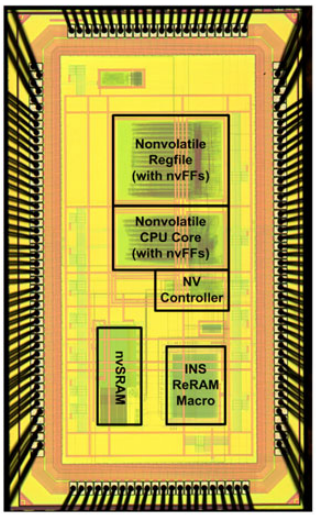


Figure 4.7.6: Measurement results of the adaptive NVP.



Affiliation	THU/NTHU
Technology	65nm SVT CMOS +ReRAM
ISA	8051
NVM	8KB ReRAM 4KB nvSRAM 1422bits nvFF
Supply	0.8V(Core), 3V(HV)
Chip Area	1560 x 2860 um^2
Nonvolatile Area	nvFFs: 5.98% nvSRAM: 4.71% ReRAM Macro: 5.84%
Fequency	>100 MHz
Active Power	33 uW/MHz@100MHz
System Restore Time	20 - 170 ns
System Restore Energy	0.45nJ (Avg)
System Store Time	4us- 1.02 ms
System Store Energy	0.40uJ (Avg)

Figure 4.7.7: Die photo and metric table.

978-1-4673-9467-3/16/$31.00